GENERALIZED LINEAR MODELS

by

Bent Jørgensen

University of Southern Denmark, Denmark

Keywords: analysis of deviance; hypothesis testing; non-normal data analysis; parameter estimation; quasi-likelihood; residual analysis; regression modeling

Abstract: Generalized linear models provide a general framework for handling regression modeling for normal and non-normal data, including multiple linear regression, ANOVA, logistic regression, Poisson regression and log-linear models for contingency tables. All the major statistical packages include facilities for fitting generalized linear models. A generalized linear model is defined by choosing a link function and a variance function, along with choosing a response variable and a set of explanatory variables. The link function transforms the mean of the response variable to a scale where the model is linear. The variance function describes how the variance behaves as a function of the mean. Each choice of variance function corresponds to a certain deviance function, and model fitting is accomplished by minimizing the deviance, generalizing least squares fitting. Inference on parameters, and hypothesis testing is performed by means of analysis of deviance, generalization the classical ANOVA method. Estimation and analysis of deviance are based on quasi-likelihood methods,

requiring only second-moment assumptions, thereby providing a certain robustness against misspecification of the probability model. The choice of link and variance functions may be checked by means of residual analysis.

Introduction

The class of *generalized linear models* was introduced in 1972 by Nelder and Wedderburn [22] as a general framework for handling a range of common statistical models for normal and non-normal data, such as multiple linear regression, ANOVA, logistic regression, Poisson regression and log-linear models. Ideas from generalized linear models are now pervasive in much of applied statistics, and are very useful in Environmetrics, where we frequently meet non-normal data, in the form of counts or skewed frequency distributions. Many common statistical packages today include facilities for fitting generalized linear models to data. Introductions to the area are given by Dobson and Barnett [8] and Firth [10], whereas Hardin and Hilbe [12] and McCullagh and Nelder [21] give more comprehensive treatments.

Suppose that we have independent data from n units $i = 1, \ldots, n$, such that for unit i we have a response variable Y_i with mean μ_i and covariates x_{ij} for $j = 1, \ldots, k$, where $x_{i1} = 1$. In ordinary multiple linear regression, the mean μ_i is assumed to be a linear function of the covariates x_{ij}, and the variance of Y_i is assumed to be common for all units. Such assumptions are seldom satisfied for non-normal data, where the linear regression model may lead to incorrect conclusions.

Generalized linear models provide a straightforward way of modeling non-normal data

when the usual regression assumptions are not satisfied. The two key ingredients for a generalized linear model are the positive *variance function* V, and the monotonic *link function* g. Both V and g are assumed to be continuously differentiable functions of the mean μ_i.

The variance of Y_i is assumed to be proportional to the variance function,

$$\text{var}(Y_i) = \sigma^2 V(\mu_i), \qquad i = 1, \ldots, n,$$

where $\sigma^2 > 0$ is the *dispersion parameter* (sometimes called the scale parameter), assumed to be common for all units. The variance function describes how the variance of the response Y_i varies as a function of the mean μ_i. The role of the link function g is to transform the mean μ_i onto a scale where the model is linear, and the regression model is hence defined by

$$g(\mu_i) = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \qquad i = 1, \ldots, n, \tag{1}$$

where $\beta_1, \ldots, \beta_k$ are unknown regression coefficients. The special case where $V(\mu) = 1$ and $g(\mu) = \mu$ recovers the assumptions of the ordinary multiple linear regression model.

Both V and g are assumed to be known functions, and may often be chosen from among a small set of standard options, reflecting basic knowledge about the nature of the response variable. Once V and g have been chosen, the regression structure is explored in much the same way as in ordinary regression or ANOVA. The analysis hence proceeds via the familiar steps of parameter estimation, model checking by residual analysis, and hypothesis testing, each of which will be discussed in more detail below.

Note that the only nonlinearity in the model (1) comes from the link function, whereas truly nonlinear models have a more complicated mean structure. However, an important

advantage of the approach is that familiar ideas from regression and ANOVA such as factors, interactions, dummy variates and polynomial regression retain their usefulness here, subject to suitable interpretations.

The Choice of Link and Variance Functions

We now present some basic guidelines for choosing the link function g and the variance function V. The role of the link function is similar to the choice of linearizing transformation traditionally used in regression and ANOVA. Rather than transforming the response variable Y, however, the link function is chosen such that the model is linear in $g(\mu)$, thereby avoiding the need for working with the mean of a transformed response variable, which may be difficult to interpret. Note that g may be selected independently of the variance function, so the question of non-constant variance is dealt with separately. Here, and in the next section, we have dropped the subscript i on Y and μ.

The choice of link and variance functions may often be guided by the nature of the domains for Y and μ. The most common choices are as follows:

- For data on the real line, where μ is a location parameter whose domain is unbounded both to the right and to the left, the *identity link* $g(\mu) = \mu$ and the *constant variance function* $V(\mu) = 1$ are commonly used, and correspond to the ordinary multiple linear regression model, including ANOVA and analysis of covariance.

- For strictly positive data, where $\mu > 0$, the *log link* $g(\mu) = \ln \mu$ is often used together with the *square variance function* $V(\mu) = \mu^2$. Possible alternatives are the *power link*

functions $g(\mu) = \mu^q$, and the *power variance functions* $V(\mu) = \mu^p$, where $q \neq 0$ and p, q are assumed to be known.

- For non-negative data, in particular counts, the *linear variance function* $V(\mu) = \mu$ is often used together with the log link, corresponding to *log-linear models*. We may also use one of the power link functions.

- For proportions satisfying $0 \leq Y \leq 1$, where $0 < \mu < 1$, a common choice is the *logit link*

$$g(\mu) = \ln \frac{\mu}{1-\mu}$$

together with the variance function $V(\mu) = \mu(1-\mu)$, which correspond to the *logistic regression model*. Other data with a bounded range, such as percentages or rating scales, may be handled in the same way after being transformed linearly onto the unit interval. Other possible link choices are the *probit link* $g(\mu) = \Phi^{-1}(\mu)$, where Φ is the standard normal CDF, and the *complementary log-log link* $g(\mu) = \ln\left[-\ln(1-\mu)\right]$.

- For a given variance function V, we define the *canonical link* g_0 as follows:

$$g_0(\mu) = \int_{\mu_0}^{\mu} \frac{1}{V(z)}\, dz, \tag{2}$$

where μ_0 is an arbitrary fixed value of the mean. Several of the link/variance function pairs proposed above have canonical link functions, see Table 1 below.

Probability Models

Up to this point have made only *second-moment assumptions*, i.e. assumptions regarding the mean and variance of the response variable. However, several of the above variance functions correspond to well-known probability models, in which case we talk about making *full distributional assumptions*.

For a given variance function V, we define the *unit deviance function* by

$$d(y;\mu) = 2\int_{\mu}^{y} \frac{y-z}{V(z)}\, dz,$$

which is strictly positive except for $y = \mu$, where it is zero. The unit deviance may be interpreted as a measure of squared distance between y and μ; in particular the case $V(\mu) = 1$ gives $d(y;\mu) = (y-\mu)^2$.

In some cases, a unit deviance function gives rise to a probability (density) function of the form

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left[-\frac{1}{2\sigma^2}d(y;\mu)\right], \tag{3}$$

where $a(y;\sigma^2)$ is a function that depends on y and the dispersion parameter σ^2 only. We call (3) an *exponential dispersion model* [14, 17]. It has mean μ and variance $\sigma^2 V(\mu)$, and we have hence obtained a probability distribution with the prescribed first and second moments. When σ^2 is known, the family (3) is called a *natural exponential family*, cf. Jørgensen [17, Ch. 2] for details.

Table 1 summarizes some common exponential dispersion models, including those that correspond to variance functions already mentioned above. For example, the constant vari-

Table 1: Summary of common exponential dispersion models (CV = coefficient of variation).

Distribution	Variance function	σ^2	Canonical link	Unit deviance
Normal	1	variance	μ	$(y-\mu)^2$
Gamma	μ^2	$(\mathrm{CV})^2$	$-1/\mu$	$2\left(\frac{y}{\mu} - \ln\frac{y}{\mu} - 1\right)$
Inverse Gaussian	μ^3	variance/μ^3	$-2/\mu^2$	$(y-\mu)^2/\left(y\mu^2\right)$
Poisson	μ	1	$\ln\mu$	$2\left(y\ln\frac{y}{\mu} + \mu - y\right)$
Binomial/m	$\mu(1-\mu)$	$1/m$	$\ln\frac{\mu}{1-\mu}$	$2\left[y\ln\frac{y}{\mu} + (1-y)\ln\frac{1-y}{1-\mu}\right]$
Negative binomial	$\mu\left(1+\mu/m\right)$	1	$\ln\frac{\mu}{m+\mu}$	$2\left[y\ln\frac{y}{\mu} + (m+y)\ln\frac{m+\mu}{m+y}\right]$

ance function $V(\mu) = 1$ gives the normal distribution with mean μ and variance σ^2. In common with the normal, the gamma distribution is a two-parameter family, where the dispersion parameter σ^2 is unknown. The three discrete distributions in Table 1 (Poisson, binomial and negative binomial) all have known values of σ^2, although the negative binomial has an additional shape parameter m. For the binomial distribution, the probability function for the proportion of success out of m trials is of the form (3) with $\sigma^2 = 1/m$.

Not every variance function and unit deviance have an associated exponential dispersion model, and this may be the case even for apparently reasonably shaped functions such as the square-root variance function $V(\mu) = \sqrt{\mu}$, see Jørgensen [17, Ch. 3]. For the above three discrete distributions, only the values of σ^2 indicated in Table 1 correspond to valid probability functions in (3), whereas in practice it is common to encounter *overdispersion,* in the form of discrete data for which σ^2 is bigger than 1 or, for binomial proportions, bigger than $1/m$.

On this background it is fortunate that most of the estimation and testing methods to be introduced below depend on second-moment assumptions for Y only, giving procedures that are robust against misspecification of the probability model, as long as the link and variance functions are correctly specified.

Parameter Estimation

We now consider estimation of the vector of regression coefficients $\boldsymbol{\beta}$ from data $\mathbf{y} = (y_1, \ldots, y_n)'$, where σ^2 is either known, or is an additional parameter to be estimated from the data. We assume that suitable link and variance functions have been chosen. Let us generalize the variance assumption as follows:

$$\text{var}(Y_i) = \frac{\sigma^2}{w_i} V(\mu_i), \qquad i = 1, \ldots, n,$$

where $w_1, \ldots, w_n$ are known weights, which may for example be sample sizes if the Y_is are group averages, where subscript i again refers to the unit. Let $\mathbf{x}_i$ denote the k-vector of covariates for unit i, let $\mathbf{X}$ be the $n \times k$ design matrix with rows $\mathbf{x}_1', \ldots, \mathbf{x}_n'$, and assume that $\mathbf{X}$ has rank $k < n$.

For binomial proportions, w_i is the number of trials for unit i and $\sigma^2 = 1$ now corresponds to the ordinary binomial distribution, whereas $\sigma^2 > 1$ indicates overdispersion.

Let us define the (total) *deviance* for $\boldsymbol{\beta}$ as the weighted sum of unit deviances,

$$D(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i d(y_i; \mu_i),$$

where here and in the following $\boldsymbol{\beta}$ enters via $\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$. We define the estimate $\widehat{\boldsymbol{\beta}}$ to

be the value of $\boldsymbol{\beta}$ that minimizes $D(\boldsymbol{\beta})$. In the case $V(\mu) = 1$, the deviance is the familiar residual sum-of-squares statistic from regression, and $\widehat{\boldsymbol{\beta}}$ is the least-squares estimate.

Under full distributional assumptions, the log likelihood for $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta}) = \text{const.} - \frac{1}{2\sigma^2} D(\boldsymbol{\beta}), \tag{4}$$

where the constant depends on σ^2 and the data only, so that $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimate. Under second-moment assumptions, (4) (without the constant) is the (log) quasi-likelihood of Wedderburn [30] and McCullagh [20], and $\widehat{\boldsymbol{\beta}}$ is the corresponding maximum quasi-likelihood estimate.

To calculate $\widehat{\boldsymbol{\beta}}$, we solve the (quasi-) score equation corresponding to (4),

$$\sum_{i=1}^{n} \mathbf{x}_i \frac{w_i \left(Y_i - \mu_i\right)}{\dot{g}(\mu_i) V(\mu_i)} = \mathbf{0}, \tag{5}$$

where $\dot{g}$ denotes the derivative of g. Note the simplification that occurs for the canonical link, where $\dot{g}(\mu_i)V(\mu_i) = 1$. The equation (5) generally is nonlinear, and is solved iteratively by Fisher's scoring method, as we shall now see.

In each step of the iteration for Fisher's scoring method, the updated value $\boldsymbol{\beta}^*$ of the regression parameter is the solution to the following weighted least-squares equation:

$$\boldsymbol{\beta}^* \mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{X}'\mathbf{W}\mathbf{z}. \tag{6}$$

Here $\mathbf{W}$ and $\mathbf{z}$, which both depend on the previous value of $\boldsymbol{\beta}$, are defined by

$$\mathbf{W} = \text{diag}\left(\frac{w_1}{\dot{g}(\mu_1)^2 V(\mu_1)}, \ldots, \frac{w_n}{\dot{g}(\mu_n)^2 V(\mu_n)}\right)$$

and

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \dot{\mathbf{g}}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}), \tag{7}$$

where $\boldsymbol{\mu}$ is the vector of means and $\dot{\mathbf{g}}(\boldsymbol{\mu}) = \text{diag}\left[\dot{g}(\mu_1), \ldots, \dot{g}(\mu_n)\right]$. The iterations are stopped when the relative decrease of the deviance becomes small.

The starting value for the iterations is obtained from the data $\mathbf{y}$, so that in the first iteration we take $\boldsymbol{\mu} = \mathbf{y}$ and replace $\mathbf{X}\boldsymbol{\beta}$ in (7) by the vector with entries $g(y_i)$, with suitable modifications for extreme values of y_i, where $g(y_i)$ and $V(y_i)$ may not be defined. By using a good weighted least-squares algorithm for solving (6), the resulting Fisher scoring algorithm becomes very efficient.

When the dispersion parameter is unknown, it may be estimated by the *Pearson Estimator,*

$$\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\frac{w_i\left(y_i - \widehat{\mu}_i\right)^2}{V(\widehat{\mu}_i)},$$

where $\widehat{\mu}_i = g^{-1}(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$ is the ith *fitted value*. Some computer packages routinely estimate σ^2 by the *deviance estimator* $D(\widehat{\boldsymbol{\beta}})/(n-k)$, but this estimator cannot be recommended in practice because of problems with bias and inconsistency in the case of a non-constant variance function. For positive data, the deviance may also be sensitive to rounding errors for small values of y_i.

The asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is estimated by the inverse (Fisher) information matrix, giving

$$\text{var}(\widehat{\boldsymbol{\beta}}) \approx \sigma^2\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}, \tag{8}$$

where $\mathbf{W}$ is calculated from $\widehat{\boldsymbol{\beta}}$. The standard error $\text{se}(\widehat{\beta}_j)$ is calculated as the square-root

of the jth diagonal element of this matrix, for $j = 1, \ldots, k$. When σ^2 is known, a $1 - \alpha$ confidence interval for β_j is defined by the endpoints

$$\widehat{\beta}_j \pm \mathrm{se}(\widehat{\beta}_j)\mathrm{z}_{1-\alpha/2}, \tag{9}$$

where $\mathrm{z}_{1-\alpha/2}$ is the $1 - \alpha/2$ standard normal quantile. For σ^2 unknown, we replace σ^2 by $\widehat{\sigma}^2$ in (8) and $\mathrm{z}_{1-\alpha/2}$ by $\mathrm{t}_{1-\alpha/2}(n-k)$ in (9), where $\mathrm{t}_{1-\alpha/2}(n-k)$ is the $1 - \alpha/2$ quantile of Student's t distribution with $n - k$ degrees of freedom.

Residual Analysis

Residuals are usually defined as observed minus fitted values, standardized to have constant variance. From this point of view, an obvious choice of residual for generalized linear models is the *Pearson residual,* defined by

$$r_{P_i} = \frac{y_i - \widehat{\mu}_i}{[V(\widehat{\mu}_i)]^{1/2}}.$$

Residuals are useful for making graphical checks of the adequacy of the link and variance function choices and other model assumptions. In order to perform such checks as accurately as possible, it is useful to work with residuals that are as nearly normally distributed as possible. Unfortunately, the Pearson residual is somewhat inadequate from this point of view, because it tends to reflect the skewness of the underlying distribution. A better choice is the *deviance residual*, defined by

$$r_{D_i} = \pm \left[d(y_i; \widehat{\mu}_i)\right]^{1/2}, \tag{10}$$

where $\pm$ denotes the sign of $y_i - \widehat{\mu}_i$. Pierce and Schafer [23] and McCullagh and Nelder [21, pp. 37–40] found that the deviance residual is much closer to being normal than the Pearson residual, but has a bias of

$$-\frac{\mathrm{E}(Y_i - \mu_i)^3}{6\sigma^2 \left[V(\mu_i)\right]^{3/2}},$$

which should be subtracted from (10). Note, however, that the bias depends on the third moment of Y_i. See also Williams [31], who studied residuals and diagnostics for generalized linear models.

Under second-moment assumptions, an alternative way of correcting the bias of the deviance residual is via the *modified deviance residual* $r^*_{D_i}$, defined by

$$r^*_{D_i} = r_{D_i} + \frac{\sigma^2}{r_{D_i}} \ln \frac{r_{W_i}}{r_{D_i}},$$

see Jørgensen [17, Ch. 3]. Here r_{W_i} is the *Wald residual,* defined by

$$r_{W_i} = \left[g_0(y_i) - g_0(\mu_i)\right] \left[V(y_i)\right]^{1/2},$$

where g_0 is the canonical link (2). Note that in the discrete case, r_{W_i} is generally infinite for extreme values of y_i.

Taking the variation of $\widehat{\mu}_i$ into account, all of the above residuals have approximately mean zero and variance $\sigma^2(1 - h_i)$, where h_i is the ith diagonal element of the *hat matrix* $\mathbf{H}$, defined by

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X} \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W}^{1/2}.$$

In practice, we hence use standardized residuals such as $r^*_{D_i}(1 - h_i)^{-1/2}$, which are nearly normal with variance σ^2. The residuals may be plotted against the fitted values (or better

their logarithms) in order to check the validity of the proposed variance function; or the correctness of the distributional assumption may be checked by means of a normal Q-Q plot for the residuals. See McCullagh and Nelder [21, Ch. 12] and Davison and Snell [5] for more details about residual analysis for generalized linear models.

Analysis of Deviance

Analysis of deviance is the method of parameter inference for generalized linear models based on the deviance, generalizing ideas from ANOVA, and first introduced by Nelder and Wedderburn [22]. We emphasize, however, that even for balanced data, the situation is similar to regression analysis, in the sense that model terms must be eliminated sequentially, and the significance of a term may depend on which other terms are in the model. We consider separately the cases where σ^2 is known and unknown, but first we introduce some notation.

Let H_1 denote the model (1) with k parameters, and let $D_1 = D(\widehat{\boldsymbol{\beta}})$ denote the minimized deviance under H_1. Similarly, let H_2 denote a sub-model of H_1 with $l < k$ parameters, and let D_2 denote the corresponding minimized deviance, where $D_2 \geq D_1$. The model H_2 may for example correspond to the hypothesis that certain regression coefficients are zero, or to some other linear constraint on $\boldsymbol{\beta}$.

The results that we now present are based on large-sample theory, and we need to consider two separate asymptotic frameworks. The first is called *large w asymptotics*, where it is assumed that the data y_i are group averages based on large sample sizes w_i for all i. This

framework is often relevant for discrete data, where the conventional rule is that all expected counts should be at least five in order for the asymptotic results to apply. For binomial data both the expected number of successes and failures should be at least five. Under full distributional assumptions, the large w asymptotics are called *small-dispersion asymptotics*, see Jørgensen [14, 15].

The second asymptotic framework is *large n asymptotics*, which is mainly relevant for regression models, where it is assumed that n is large relative to the number of parameters in the models under consideration.

Known dispersion parameter

The case of a known dispersion parameter σ^2 is mainly relevant for discrete data, as discussed in connection with Table 1. We assume for simplicity that $\sigma^2 = 1$.

The deviance D_1 is a measure of goodness-of-fit of the model H_1, and is also known as the G^2 *statistic* in discrete data analysis [1, p. 48]. A more traditional goodness-of-fit statistic is the *Pearson X^2 statistic*

$$X^2 = \sum_{i=1}^{n} \frac{w_i \left(y_i - \widehat{\mu}_i\right)^2}{V(\widehat{\mu}_i)},$$

which we have already met in connection with the Pearson estimator above.

Asymptotically for large w, the statistics D_1 and X^2 are equivalent and distributed as $\chi^2(n-k)$ under H_1, but various numerical and analytical investigations have shown that the limiting χ^2 distribution is approached faster for the X^2 statistic than for D_1, at least for discrete data [4], in line with our recommendation of the Pearson estimator for σ^2 above. A

formal level α goodness-of-fit test for H_1 is obtained by rejecting H_1 if $X^2 > \chi^2_{1-\alpha}(n-k)$, the latter being the $1-\alpha$ quantile of the $\chi^2(n-k)$ distribution. This test may be interpreted as a test for overdispersion [6]. Here and in the following, we may calculate the P value of the test in the usual way; in the present case by equating X^2 to $\chi^2_{1-P}(n-k)$ and solving for P.

There is a long tradition for goodness-of-fit tests for discrete data, but it should be kept in mind that the fit of a model is a complex question, which can hardly be summarized in a single number. For this reason, we recommend that the X^2 test is supplemented with an inspection of residuals, as discussed above.

Once H_1 has been accepted, we may calculate confidence intervals for the regression coefficients using the normal distribution, as explained in the parameter estimation section. We may also proceed to test the sub-model H_2 under H_1. For this purpose we use the log (quasi-) likelihood ratio statistic $D_2 - D_1$, which is a relative measure of fit for H_2 under H_1. The asymptotic distribution of $D_2 - D_1$ is $\chi^2(k-l)$ for n large as well as for w large, and H_2 is rejected at level α if $D_2 - D_1 > \chi^2_{1-\alpha}(k-l)$. Once H_2 has been accepted, a sub-model of H_2 may be tested under H_2 in a similar way, and so on.

In the case where the known value of σ^2 is different from 1, we use the *scaled deviance* D_1/σ^2 instead of D_1, and the *scaled Pearson statistic* X^2/σ^2 instead of X^2 and so on.

Unknown dispersion parameter

The dispersion parameter is usually unknown for continuous data, as discussed in connection with Table 1, and the methods of inference need to be modified accordingly. In the discrete case we may prefer to work with unknown dispersion parameter, if evidence of overdispersion has been found in the data.

When the dispersion parameter is unknown, there is no formal goodness-of-fit test available based on X^2 as above. Instead, X^2 is used for estimating the dispersion parameter, as explained in the parameter estimation section, and the fit of the model H_1 to the data must be checked by residual analysis.

Once the fit of H_1 has been verified, we may set up confidence intervals for the regression coefficients using the t distribution, as explained in the parameter estimation section. Similarly, we may use the large w asymptotic $\chi^2(n-k)$ distribution for $(n-k)\widehat{\sigma}^2/\sigma^2$ to calculate confidence intervals for σ^2.

Let us now consider testing H_2 under H_1 for σ^2 unknown. A simple-minded approach is to base the test on the scaled deviance difference $\Delta = (D_2 - D_1)/\widehat{\sigma}^2$, whose large n asymptotic distribution under H_2 is $\chi^2(k-l)$, due to the consistency of the dispersion parameter estimate $\widehat{\sigma}^2$ in this limit. Contrary to the case σ^2 known, however, the asymptotic χ^2 distribution for Δ does not apply in the large w limit. Instead, we scale Δ by the degrees of freedom to obtain the following F statistic:

$$F = \frac{D_2 - D_1}{(k-l)\widehat{\sigma}^2},$$

whose asymptotic distribution is $\mathrm{F}(k-l, n-k)$ for w large, which agrees asymptotically

with the limiting $\chi^2(k-l)$ distribution for Δ in the large n limit. The F test is hence valid in both the large w and large n limits, and we reject H_2 at level α if $F > \mathrm{F}_{1-\alpha}(k-l, n-k)$. We may proceed similarly to test, in a sequential manner, further reductions of the model.

The t based confidence intervals for the regression coefficients mentioned above may be justified by similar arguments. Further details regarding the asymptotic results may be found in Jørgensen [14].

Generalizations

Generalized linear models have now been extended in many different directions compared with Nelder and Wedderburn's original definition. The ideas of quasi-likelihood and estimating functions have made it easy to develop simple and robust estimation methods for a wide variety of problems, including correlated data, while preserving much of the original simplicity of the idea.

In particular, Liang and Zeger [19] proposed the method of *generalized estimating equations* (GEE) for analysis of longitudinal data, which is now widely used, and has spawned much further research [7]. Several methods for analysis of generalized linear mixed models have been proposed, see Schall [26], Zeger and Karim [32], Breslow and Clayton [3] and Lee and Nelder [18].

Efron [9] and Smyth [27] proposed methods for generalized linear models where the dispersion parameter, as well as the mean, varies as a function of covariates according to a specified regression model.

A large variety of probability models are of exponential dispersion model form, and this class has been studied extensively together with the even larger class of *dispersion models*, see Jørgensen [17]. The generalization of analysis of deviance to dispersion models was investigated by Jørgensen [13, 16].

Software

The GLIM (Generalized Linear Interactive Modelling) software, specially designed for fitting generalized linear models, was first released in 1974, and went through several releases, the last one being GLIM4 from 1993. See also the monograph [2] on statistical modeling in GLIM4. GLIM is no longer available on the market, but all the major statistical packages now include facilities for fitting generalized linear models, including GenStat [11], R [24], SAS [25], S-Plus [28] and STATA [29]. These implementations of generalized linear models basically retain the simplicity and flexibility that characterized the original implementation in GLIM, allowing the user to select any suitable combination of link function and variance function, combined with the Wilkinson and Rogers notation for specifying the regression model.

Acknowledgements

I am grateful to Clarice G.B. Demétrio for useful comments in relation to the paper. This research was supported by the Danish Natural Science Research Council.

References

[1] Agresti, A. (2002). *Categorical Data Analysis,2nd Ed.* Wiley, New York.

[2] Aitkin, M., Francis, B. & Hinde, J. (2005). *Statistical Modelling in GLIM4, 2nd Ed.* Oxford University Press, New York.

[3] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

[4] Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B* **46**, 440–464.

[5] Davison, A.C. & Snell, E.J. (1991). Residuals and diagnostics. In *Statistical Theory and Modelling,* D.V. Hinkley, N. Reid, and E.J. Snell, eds. Chapman & Hall, London.

[6] Dean, C.B. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association* **87**, 451–463.

[7] Diggle, P.J., Heagerty, P., Liang, K.-Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data,2nd Ed.* Oxford University Press, Oxford.

[8] Dobson, A.J. & Barnett, A.G. (2008). *An Introduction to Generalized Linear Models, 3rd Ed.* Chapman & Hall/CRC, Boca Raton.

[9] Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* **81**, 709–721.

[10] Firth, D. (1991). Generalized linear models. In *Statistical Theory and Modelling,* D.V. Hinkley, N. Reid, and E.J. Snell, eds. Chapman & Hall, London.

[11] GenStat 13th Edition (2011). VSN International Ltd., Hemel Hempstead, UK.

[12] Hardin, J.W. and Hilbe, J.M. (2007). *Generalized Linear Models and Extensions, 2nd Ed.* Stata Press, College Station.

[13] Jørgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70**, 19–28.

[14] Jørgensen, B. (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society Series B* **49**, 127–162.

[15] Jørgensen, B. (1987). Small-dispersion asymptotics. *Brazilian Journal of Probability and Statistics* **1**, 59–90.

[16] Jørgensen, B. (1992). Exponential dispersion models and extensions: A review. *International Statistical Review* **60**, 5–20.

[17] Jørgensen, B. (1997). *The Theory of Dispersion Models.* Chapman & Hall, London.

[18] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society Series B* **58**, 619–678.

[19] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

[20] McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics* **11**, 59–67.

[21] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models, 2nd Ed.* Chapman & Hall, London.

[22] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A* **135**, 370–384.

[23] Pierce, D.A. & Schafer, D.W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association* **81**, 977–986.

[24] R Development Core Team (2010). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[25] SAS Version 9.2 (2011). SAS Institute, Cary, North Carolina.

[26] Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–727.

[27] Smyth, G.K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society Series B* **51**, 47–60.

[28] S-Plus Version 8.2 (2011). TIBCO Software Inc. Palo Alto, California.

[29] STATA Version 11 (2011). StataCorp LP, College Station, Texas.

[30] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

[31] Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* **36**, 181–191. (Residuals and diagnostics for generalized linear models.)

[32] Zeger, S.L. & Karim, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.

Bibliography

Breslow, N.E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44. (Overdispersion in log-linear Poisson models.)

Crawley, M.J. (2007). *The R Book.* Wiley, Chichester. (General reference work on statistical analysis using R. Includes several chapters on generalized linear and related models.)

Czado, C. (1994). Parametric link modification of both tails in binary regression. *Statistical Papers* **35**, 189–201. (Parametric link functions for logistic regression.)

Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag, New York. (Generalized linear models for multivariate data.)

Faraway, J.J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Chapman & Hall/CRC, Boca Raton.

(Statistical textbook with emphasis on R. Includes several chapters on generalized linear and related models.)

Gill, J. (2001). *Generalized linear models. A Unified Approach.* Sage Publications, Thousand Oaks. (Statistical textbook on generalized linear models for the social sciences.)

Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman & Hall, London. (Generalized linear models with the regression part replaced by a sum of smooth functions.)

Jørgensen, B. (1997). Proper dispersion models (with discussion). *Brazilian Journal of Probability and Statistics* **11**, 89–140. (Introduces proper dispersion models, and studies estimation and testing for generalized linear models with proper dispersion model errors.)

Jørgensen, B. & Souza, M.P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal 1994,* 69–93. (Analysis of Brazilian car insurance data using generalized linear models; the technique developed is suitable for non-negative continuous data with a positive probability mass in zero.)

Lindsey, J.K. (1997). *Applying Generalized Linear Models, Corrected 3rd printing.* Springer-Verlag, Heidelberg. (Statistical textbook that describes how generalized linear modelling procedures can be used in many different fields, such as survival models, time series, and spatial analysis.)

Madsen, H. and Thyregod, P. (2011). *Introduction to General and Generalized Linear Models.* Chapman & Hall/CRC, Boca Raton. (Statistical textbook that covers both linear and generalized linear models from a likelihood point of view.)

Myers, R.H., Montgomery, D.C., Vining, G.G. and Robinson, T.J. (2010). *Generalized linear models With Applications in Engineering and the Sciences, 2nd Ed.* Wiley, Hoboken. (Statistical textbook that covers linear and nonlinear regression models, generalized linear models, generalized estimating equations, and generalized linear mixed models.)

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics* **29**, 15–24. (Diagnostics based on one-parameter link families.)

Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics* **31**, 144–148. (Overdispersion in logistic regression.)