CHAPTER 2

An outline of generalized linear models

2.1 Processes in model fitting

In Chapter 1 we considered briefly some of the reasons for model fitting as an aid for interpreting data. Before describing the form of generalized linear models (GLMs) we look first at the processes of model fitting, following closely the ideas of Box and Jenkins (1976), which they applied to time series. Three processes are distinguished: (i) model selection, (ii) parameter estimation and (iii) prediction of future values. Box and Jenkins use 'model identification' in place of our 'model selection', but we prefer to avoid any implication that a correct model can ever be known with certainty. In distinguishing these three processes, we do not assume that an analysis consists of the successive application of each just once. In practice there are backward steps, false assumptions that have to be retracted, and so on.

We now look briefly at some of the ideas associated with each of the three processes.

2.1.1 *Model selection*

Models that we select to fit to data are usually chosen from a particular class and, if the model-fitting process is to be useful, this class must be broadly relevant to the kind of data under study. An important characteristic of generalized linear models is that they assume independent (or at least uncorrelated) observations. More generally, the observations may be independent in blocks of fixed known sizes. As a consequence, data exhibiting the autocorrelations of time series and spatial processes are expressly excluded. This assumption of independence is characteristic of the linear models of classical regression analysis, and is carried over without modification to the wider class of generalized linear models. In Chapter 9 we look at the possibility of relaxing this assumption. A second assumption about the error structure is that there is a single error term in the model. This constraint excludes, for instance, models for the analysis of experiments having more than one explicit error term. Perhaps the simplest instance of a model excluded by this criterion is the standard linear model for the split-plot design, which has two error terms, one for between-whole-plot variance and one for within-whole-plot variance. Again, we shall later relax this restriction for certain kinds of GLMs.

In practice, these two restrictions on the form of the error distribution are less restrictive than they might appear at first sight. For instance autoregressive models can easily be fitted using programmes designed expressly for ordinary linear models. Further, certain forms of dependence, such as that occurring in the analysis of contingency tables where a certain marginal total is fixed, can in fact be handled as if the observations were independent. Similarly, though a grouping factor corresponding to a nuisance classification may induce correlations within groups, a within-groups analysis after elimination of the effects of that nuisance factor can proceed as if the observations were independent.

The choice of scale for analysis is an important aspect of model selection. A common choice is between an analysis of Y, i.e. the original scale, or $\log Y$. To the question 'What characterizes a "good" scale?' we must answer that it all depends on the purpose for which the scale is to be used. To quote from the preface to the first edition in Jeffreys (1961): 'It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude'. In classical linear regression analysis a good scale should combine constancy of variance, approximate Normality of errors and additivity of systematic effects. Now there is usually no *a priori* reason to believe that such a scale exists, and it is not difficult to imagine cases in which it does not. For instance, in the analysis of discrete data where the errors are well approximated by the Poisson distribution, the systematic effects are often multiplicative. Here $Y^{1/2}$ gives approximate constancy of variance, $Y^{2/3}$ does better for approximate symmetry or Normality, and $\log Y$ produces additivity of the systematic effects. Evidently, no single scale will simultaneously produce all the desired properties.

With the introduction of generalized linear models, scaling problems are greatly reduced. Normality and constancy of variance are no longer required, although the way in which the variance depends on the mean must be known. Additivity of effects, while still an important component of all generalized linear models, can be specified to hold on a transformed scale if necessary. In generalized linear models, additivity is, correctly, postulated as a property of the expected responses. Additivity with respect to the data themselves can only ever be a rough approximation.

There remains the problem in model selection of the choice of x-variables (or covariates as we shall call them) to be included in the systematic part of the model. There is a large literature on this topic in linear models. In its simplest form, we are given a number of candidate covariates, $x_1, \ldots, x_p$, and are required to find a subset of these that is in some sense best for constructing the fitted values

$$\hat{\boldsymbol{\mu}} = \sum \mathbf{x}_j \hat{\beta}_j.$$

Implicit in the strategies that have been proposed is that there is a balance to be struck between improving the fit to the observed data by adding an extra term to the model and the usually undesirable increase in complexity implicit in this extra term. Note that even if we could define exactly what is meant by an optimum model in a given context, it is most unlikely that the data would indicate a clear winner among the potentially large number of competing models. We must anticipate that, clustered around the 'best' model will be a set of alternatives almost as good and not statistically distinguishable. Selection of covariates is discussed at various points in the chapters that follow, particularly in section 3.9 and in the various examples.

2.1.2 *Estimation*

Having selected a particular model, it is required to estimate the parameters and to assess the precision of the estimates. In the case of generalized linear models, estimation proceeds by defining a measure of goodness of fit between the observed data and the fitted values generated by the model. The parameter estimates are the values that minimize the goodness-of-fit criterion. We shall be concerned primarily with estimates obtained by maximizing the likelihood or log likelihood of the parameters for the data observed. If $f(y;\theta)$ is the density function or probability distribution for the observation y given the parameter θ, then the log likelihood, expressed as a function of the mean-value parameter, $\mu = E(Y)$, is just

$$l(\mu;y) = \log f(y;\theta).$$

The log likelihood based on a set of independent observations $y_1, \ldots, y_n$ is just the sum of the individual contributions, so that

$$l(\boldsymbol{\mu};\mathbf{y}) = \sum_i \log f_i(y_i;\theta_i)$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Note that the density function $f(y;\theta)$ is considered as a function of y for fixed θ whereas the log likelihood is considered primarily as a function of θ for the particular data y observed. Hence the reversal of the order of the arguments.

There are advantages in using as the goodness-of-fit criterion, not the log likelihood $l(\boldsymbol{\mu};\mathbf{y})$ but a particular linear function, namely

$$D^*(\mathbf{y};\boldsymbol{\mu}) = 2l(\mathbf{y};\mathbf{y}) - 2l(\boldsymbol{\mu};\mathbf{y}),$$

which we call the *scaled deviance.* Note that, for the exponential-family models considered here, $l(\mathbf{y};\mathbf{y})$ is the maximum likelihood achievable for an exact fit in which the fitted values are equal to the observed data. Because $l(\mathbf{y};\mathbf{y})$ does not depend on the parameters, maximizing $l(\boldsymbol{\mu};\mathbf{y})$ is equivalent to minimizing $D^*(\mathbf{y};\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$, subject to the constraints imposed by the model.

For Normal-theory linear regression models with known variance σ^2, we have for a single observation

$$f(y;\mu) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right),$$

so that the log likelihood is

$$l(\mu;y) = -\tfrac{1}{2}\log(2\pi\sigma^2) - (y-\mu)^2/(2\sigma^2).$$

Setting $\mu = y$ gives the maximum achievable log likelihood, namely

$$l(y;y) = -\tfrac{1}{2}\log(2\pi\sigma^2),$$

so that the scaled deviance function is

$$D^*(y;\mu) = 2\{l(y;y) - l(\mu;y)\} = (y-\mu)^2/\sigma^2.$$

Apart, therefore, from the known factor σ^2, the deviance in this instance is identical to the residual sum of squares and minimum deviance is synonymous with least squares.

2.1.3 *Prediction*

Prediction, as interpreted here, is concerned with answers to 'what-if' questions of the kind that may be posed following a statistical analysis. In the context of a time series such a question might take the form 'what is the predicted value of the response at time t in the future, given the past history of the series and the model used in the analysis?'. More generally, prediction is concerned with statements about the likely values of unobserved events, not necessarily those in the future. For example, following an analysis of the incidence of heart disease nationally, the data being classified by region and age-group, a typical 'what-if' question is 'what would be the predicted incidence for a particular city if it had the same age structure as the country as a whole?'. This kind of prediction is an instance of standardization. For another example, consider a quantal response assay in which we measure the proportion of subjects responding to a range of dose levels. We fit a model expressing how this proportion varies with dose, and from the fitted model we predict the dose that gives rise to a 50% response rate, the so-called LD50. This answers the question 'what would be the predicted dose if the response rate were 50%?'. The word *calibration* is often used here to distinguish inverse prediction problems, in which the response is fixed and we are required to make statements about the likely values of x, from the more usual type in which the roles are reversed.

To be useful, predicted quantities need to be accompanied by measures of precision. These are ordinarily calculated on the assumption that the set-up that produced the data remains constant, and that the model used in the analysis is substantially correct. For an account of prediction as a unifying idea connecting the analysis of covariance and various kinds of standardization, see Lane and Nelder (1982).

2.2 The components of a generalized linear model

Generalized linear models are an extension of classical linear models, so that the latter form a suitable starting point for discussion. A vector of observations $\mathbf{y}$ having n components is assumed to be a realization of a random variable $\mathbf{Y}$ whose components are independently distributed with means $\boldsymbol{\mu}$. The systematic part of the model is a specification for the vector $\boldsymbol{\mu}$ in terms of a small number of unknown parameters $\beta_1, \ldots, \beta_p$. In the case of ordinary linear models, this specification takes the form

$$\boldsymbol{\mu} = \sum_1^p \mathbf{x}_j \beta_j, \tag{2.1}$$

where the βs are parameters whose values are usually unknown and have to be estimated from the data. If we let i index the observations then the systematic part of the model may be written

$$E(Y_i) = \mu_i = \sum_1^p x_{ij}\beta_j; \qquad i = 1, \ldots, n, \tag{2.2}$$

where x_{ij} is the value of the jth covariate for observation i. In matrix notation (where $\boldsymbol{\mu}$ is $n \times 1$, $\mathbf{X}$ is $n \times p$ and $\boldsymbol{\beta}$ is $p \times 1$) we may write

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X}$ is the model matrix and $\boldsymbol{\beta}$ is the vector of parameters. This completes the specification of the systematic part of the model.

For the random part we assume independence and constant variance of errors. These assumptions are strong and need checking, as far as is possible, from the data themselves. We shall consider techniques for doing this in Chapter 12. Similarly, the structure of the systematic part assumes that we know the covariates that influence the mean and can measure them effectively without error; this assumption also needs checking, as far as is possible.

A further specialization of the model involves the stronger assumption that the errors follow a Gaussian or Normal distribution with constant variance σ^2.

We may thus summarize the classical linear model in the form:

The components of $\mathbf{Y}$ are independent Normal variables with constant variance σ^2 and

$$\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu} \quad \text{where} \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \tag{2.3}$$

2.2.1 *The generalization*

To simplify the transition to generalized linear models, we shall rearrange (2.3) slightly to produce the following three-part specification:

1. The *random component*: the components of $\mathbf{Y}$ have independent Normal distributions with $E(\mathbf{Y}) = \boldsymbol{\mu}$ and constant variance σ^2;
2. The *systematic component*: covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ produce a *linear predictor* $\boldsymbol{\eta}$ given by

$$\boldsymbol{\eta} = \sum_1^p \mathbf{x}_j \beta_j;$$

3. The *link* between the random and systematic components:

$$\boldsymbol{\mu} = \boldsymbol{\eta}.$$

This generalization introduces a new symbol η for the linear predictor and the third component then specifies that μ and η are in fact identical. If we write

$$\eta_i = g(\mu_i),$$

then $g(\cdot)$ will be called the *link function*. In this formulation, classical linear models have a Normal (or Gaussian) distribution in component 1 and the identity function for the link in component 3. Generalized linear models allow two extensions; first the distribution in component 1 may come from an exponential family other than the Normal, and secondly the link function in component 3 may become any monotonic differentiable function.

We look first at the extended distributional assumption.

2.2.2 *Likelihood functions for generalized linear models*

We assume that each component of $\mathbf{Y}$ has a distribution in the exponential family, taking the form

$$f_Y(y;\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\} \qquad (2.4)$$

for some specific functions $a(\cdot), b(\cdot)$ and $c(\cdot)$. If ϕ is known, this is an exponential-family model with canonical parameter θ. It may or may not be a two-parameter exponential family if ϕ is unknown. Thus for the Normal distribution

$$\begin{aligned} f_Y(y;\theta,\phi) &= \frac{1}{\sqrt{(2\pi\sigma^2)}}\exp\{-(y-\mu)^2/2\sigma^2\} \\ &= \exp\{(y\mu - \mu^2/2)/\sigma^2 - \tfrac{1}{2}(y^2/\sigma^2 + \log(2\pi\sigma^2))\}, \end{aligned}$$

so that $\theta = \mu$, $\phi = \sigma^2$, and

$$a(\phi) = \phi, \quad b(\theta) = \theta^2/2, \quad c(y,\phi) = -\tfrac{1}{2}\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}.$$

We write $l(\theta,\phi;y) = \log f_Y(y;\theta,\phi)$ for the log-likelihood function considered as a function of θ and ϕ, y being given. The mean and variance of Y can be derived easily from the well known relations

$$E\left(\frac{\partial l}{\partial\theta}\right) = 0 \qquad (2.5)$$

and

$$E\left(\frac{\partial^2 l}{\partial\theta^2}\right) + E\left(\frac{\partial l}{\partial\theta}\right)^2 = 0. \qquad (2.6)$$

We have from (2.4) that

$$l(\theta;y) = \{y\theta - b(\theta)\}/a(\phi) + c(y,\phi),$$

whence

$$\frac{\partial l}{\partial\theta} = \{y - b'(\theta)\}/a(\phi) \qquad (2.7)$$

and

$$\frac{\partial^2 l}{\partial\theta^2} = -b''(\theta)/a(\phi), \qquad (2.8)$$

where primes denote differentiation with respect to θ.

From (2.5) and (2.7) we have

$$0 = E\left(\frac{\partial l}{\partial \theta}\right) = \{\mu - b'(\theta)\}/a(\phi),$$

so that

$$E(Y) = \mu = b'(\theta).$$

Similarly from (2.6), (2.7) and (2.8) we have

$$0 = -\frac{b''(\theta)}{a(\phi)} + \frac{\text{var}(Y)}{a^2(\phi)},$$

so that

$$\text{var}(Y) = b''(\theta)a(\phi).$$

Thus the variance of Y is the product of two functions; one, $b''(\theta)$, depends on the canonical parameter (and hence on the mean) only and will be called the *variance function*, while the other is independent of θ and depends only on ϕ. The variance function considered as a function of μ will be written $V(\mu)$.

The function $a(\phi)$ is commonly of the form

$$a(\phi) = \phi/w,$$

where ϕ, also denoted by σ^2 and called the *dispersion parameter*, is constant over observations, and w is a known *prior weight* that varies from observation to observation. Thus for a Normal model in which each observation is the mean of m independent readings we have

$$a(\phi) = \sigma^2/m,$$

so that $w = m$.

The most important distributions of the form (2.4) with which we shall be concerned are summarized in Table 2.1.

Table 2.1 *Characteristics of some common univariate distributions in the exponential family*[†]

	Normal	*Poisson*	*Binomial*	*Gamma*	*Inverse Gaussian*
Notation	$N(\mu,\sigma^2)$	$P(\mu)$	$B(m,\pi)/m$	$G(\mu,\nu)$	$IG(\mu,\sigma^2)$
Range of y	$(-\infty,\infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0,\infty)$	$(0,\infty)$
Dispersion parameter: ϕ	$\phi=\sigma^2$	1	$1/m$	$\phi=\nu^{-1}$	$\phi=\sigma^2$
Cumulant function: $b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1+e^\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y;\phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$	$-\log y!$	$\log\binom{m}{my}$	$\nu\log(\nu y)-\log y$ $-\log\Gamma(\nu)$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3)+\frac{1}{\phi y}\right\}$
$\mu(\theta)=E(Y;\theta)$	θ	$\exp(\theta)$	$e^\theta/(1+e^\theta)$	$-1/\theta$	$(-2\theta)^{-1/2}$
Canonical link: $\theta(\mu)$	identity	log	logit	reciprocal	$1/\mu^2$
Variance function: $V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

[†] The mean-value parameter is denoted by μ, or by π for the binomial distribution.
The parameterization of the gamma distribution is such that its variance is μ^2/ν.
The canonical parameter, denoted by θ, is defined by (2.4). The relationship between μ and θ is given in lines 6 and 7 of the Table.

2.2.3 *Link functions*

The link function relates the linear predictor η to the expected value μ of a datum y. In classical linear models the mean and the linear predictor are identical, and the identity link is plausible in that both η and μ can take any value on the real line. However, when we are dealing with counts and the distribution is Poisson, we must have $\mu > 0$, so that the identity link is less attractive, in part because η may be negative while μ must not be. Models for counts based on independence in cross-classified data lead naturally to multiplicative effects, and this is expressed by the log link, $\eta = \log \mu$, with its inverse $\mu = e^{\eta}$. Now additive effects contributing to η become multiplicative effects contributing to μ and μ is necessarily positive.

For the binomial distribution we have $0 < \mu < 1$ and a link should satisfy the condition that it maps the interval $(0, 1)$ on to the whole real line. We shall consider three principal functions in subsequent chapters, namely:

1. *logit*
$$\eta = \log\{\mu/(1-\mu)\};$$
2. *probit*
$$\eta = \Phi^{-1}(\mu);$$
where $\Phi(\cdot)$ is the Normal cumulative distribution function;
3. *complementary log-log*
$$\eta = \log\{-\log(1-\mu)\}.$$

The power family of links is important, at least for observations with positive mean. This family can be specified either by

$$\eta = (\mu^{\lambda} - 1)/\lambda \qquad (2.9a)$$

with the limiting value

$$\eta = \log \mu; \quad \text{as} \quad \lambda \rightarrow 0, \qquad (2.9b)$$

or by

$$\eta = \begin{cases} \mu^{\lambda}; & \lambda \neq 0, \\ \log \mu; & \lambda = 0. \end{cases} \qquad (2.10)$$

The first form has the advantage of a smooth transition as λ passes through zero, but with either form special action has to be taken in any computation with $\lambda = 0$.

2.2.4 *Sufficient statistics and canonical links*

Each of the distributions in Table 2.1 has a special link function for which there exists a sufficient statistic equal in dimension to $\boldsymbol{\beta}$ in the linear predictor $\boldsymbol{\eta} = \sum \mathbf{x}_j \beta_j$. These *canonical links*, as they will be called, occur when

$$\theta = \eta,$$

where θ is the canonical parameter as defined in (2.4) and shown in Table 2.1. The canonical links for the distributions in that table are thus:

Normal	$\eta = \mu,$
Poisson	$\eta = \log \mu,$
binomial	$\eta = \log\{\pi/(1-\pi)\},$
gamma	$\eta = \mu^{-1},$
inverse Gaussian	$\eta = \mu^{-2}.$

For the canonical links, the sufficient statistic is $\mathbf{X}^T\mathbf{Y}$ in vector notation, with components

$$\sum_i x_{ij} Y_i, \qquad j = 1, \ldots, p,$$

summation being over the units. Note however, that, although the canonical links lead to desirable statistical properties of the model, particularly in small samples, there is in general no a priori reason why the systematic effects in a model should be additive on the scale given by that link. It is convenient if they are, but convenience alone must not replace quality of fit as a model selection criterion. In later chapters we shall deal with several models in which non-canonical links are used. We shall find, however, that the canonical links are often eminently sensible on scientific grounds.

2.3 Measuring the goodness of fit

2.3.1 *The discrepancy of a fit*

The process of fitting a model to data may be regarded as a way of replacing a set of data values $\mathbf{y}$ by a set of fitted values $\hat{\boldsymbol{\mu}}$ derived from a model involving usually a relatively small number of parameters. In general the μs will not equal the ys exactly, and the question then arises of how discrepant they are, because while a small discrepancy might be tolerable a large discrepancy is not. Measures of discrepancy or goodness of fit may be formed in various ways, but we shall be primarily concerned with that formed from the logarithm of a ratio of likelihoods, to be called the *deviance*.

Given n observations we can fit models to them containing up to n parameters. The simplest model, the *null model*, has one parameter, representing a common μ for all the ys; the null model thus consigns all the variation between the ys to the random component. At the other extreme the *full model* has n parameters, one per observation, and the μs derived from it match the data exactly. The full model thus consigns all the variation in the ys to the systematic component leaving none for the random component.

In practice the null model is usually too simple and the full model is uninformative because it does not summarize the data but merely repeats them in full. However, the full model gives us a baseline for measuring the discrepancy for an intermediate model with p parameters.

It is convenient to express the log likelihood in terms of the mean-value parameter $\boldsymbol{\mu}$ rather than the canonical parameter $\boldsymbol{\theta}$. Let $l(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})$ be the log likelihood maximized over $\boldsymbol{\beta}$ for a fixed value of the dispersion parameter ϕ. The maximum likelihood achievable in a full model with n parameters is $l(\mathbf{y}, \phi; \mathbf{y})$, which is ordinarily finite. The discrepancy of a fit is proportional to twice the difference between the maximum log likelihood achievable and that achieved by the model under investigation. If we denote by $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\mu}})$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{y})$ the estimates of the canonical parameters under the two models, the discrepancy, assuming $a_i(\phi) = \phi/w_i$, can be written

$$\sum 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}/\phi = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi,$$

where $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is known as the *deviance* for the current model and is a function of the data only. Note that

$$D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = D(\mathbf{y};\hat{\boldsymbol{\mu}})/\phi,$$

so that the scaled deviance $D^*(\mathbf{y};\boldsymbol{\mu})$ as defined in section 2.1.2 is the deviance expressed as a multiple of the dispersion parameter.

The forms of the deviances for the distributions given in Table 2.1 are as follows, summation being over $i = 1, \ldots, n$:

Normal	$\sum(y-\hat{\mu})^2,$
Poisson	$2\sum\{y\log(y/\hat{\mu}) - (y-\hat{\mu})\},$
binomial	$2\sum\{y\log(y/\hat{\mu}) + (m-y)\log[(m-y)/(m-\hat{\mu})]\},$
gamma	$2\sum\{-\log(y/\hat{\mu}) + (y-\hat{\mu})/\hat{\mu}\},$
inverse Gaussian	$\sum(y-\hat{\mu})^2/(\hat{\mu}^2 y).$

For the Normal distribution the deviance is just the residual sum of squares, while for the Poisson it is the statistic labelled G^2 by Bishop, Fienberg and Holland (1975) and others. The second term in the expressions for the Poisson and gamma deviances is often omitted for brevity. Provided that the fitted model includes a constant term, or intercept, the sum over the units of the second term is identically zero, justifying its omission. For details, see Nelder and Wedderburn, (1972).

The other important measure of discrepancy is the generalized Pearson X^2 statistic, which takes the form

$$X^2 = \sum(y-\hat{\mu})^2/V(\hat{\mu}),$$

where $V(\hat{\mu})$ is the estimated variance function for the distribution concerned. For the Normal distribution, X^2 is again the residual sum of squares, while for the Poisson or binomial distributions it is the original Pearson X^2 statistic.

Both the deviance and the generalized Pearson X^2 have exact χ^2 distributions for Normal-theory linear models (assuming of course that the model is true), and asymptotic results are available for the other distributions. However, asymptotic results may not be specially relevant to statistics calculated from limited amounts of data, and for these either D or X^2 may prove superior in its distributional properties. The deviance has a general advantage as a measure of discrepancy in that it is additive for nested sets of models if maximum-likelihood estimates are used, whereas X^2 in general is not. However, X^2 may sometimes be preferred because of its more direct interpretation.

2.3.2 *The analysis of deviance*

The analysis of variance, particularly when applied to orthogonal data with Normal errors, is a highly useful technique for screening the effects of factors and their interactions. We need some generalization of it applicable to the wider class of generalized linear models. There are two aspects of the generalization that need consideration: first, the terms in the model will, in general, no longer be orthogonal and secondly, sums of squares will, for non-Normal distributions, no longer be appropriate measures of the contribution of a term to the total discrepancy. The second problem is the more easily dealt with, and we consider it first. The terms in the analysis of variance can usefully be thought of as the first differences of the goodness-of-fit statistic for a sequence of models, each including one term more than the previous one. Thus the factorial model for two factors A and B gives rise to an analysis of variance with three terms A, B and the interaction $A.B$. The sums of squares for these are the first differences of the residual sums of squares obtained from fitting successively the models 1, A, $A+B$ and $A+B+A.B$, where 1 stands for the null model containing only the intercept. As an example, consider the following analysis of an unreplicated 4 × 3 factorial design indexed by A and B:

			Analysis of variance		
Model	*d.f.*	*Discrepancy*	*s.s.*	*d.f.*	*Term*
1	11	1000			
			500	3	A ignoring B
A	8	500			
			300	2	B eliminating A
$A+B$	6	200			
			200	6	$A.B$ eliminating A and B
$A+B+A.B$	0	0			

On the left is the sequence of models with their discrepancies, as measured by the residual sums of squares; note that the discrepancy for model 1 is just the total sum of squares about the mean in the analysis-of-variance table, while the last model is the full model, i.e. has as many parameters as observations, so that the degrees of freedom (d.f.) and the discrepancy are both zero. On the right is the analysis-of-variance table, with the sums of squares (s.s.) obtained from the first differences of the discrepancies.

The form of the generalization is now clear. Given a sequence of nested models we can use the deviance as our generalized measure of discrepancy, and form an analysis-of-deviance table by taking the first differences, as before. However, the interpretation of this table is now complicated by the non-orthogonality of the terms. Each number represents the variation accounted for by its corresponding term having eliminated the effects of those terms above it, but ignoring any effects of those terms below it. We may thus need to consider several model sequences, each producing its own analysis-of-deviance table. Note that this problem is present with classical linear models when non-orthogonality occurs. We shall not discuss here the various strategies that have been proposed for generating and looking at the goodness of fit of sets of model sequences. Suffice it to say that the aim of these strategies is to produce parsimonious models for the data in which terms that are not necessary are excluded. Note the use of the plural in 'models'; it is most unlikely with complex data that a single model will be a clear winner, and it can be most misleading to quote only the 'best' model, when several others are very close to it in terms of goodness of fit.

Once we depart from the Normal-theory linear model we generally lack an exact theory for the distribution of the deviance. In certain special cases, for example with observations in a simple design having exponential or inverse Gaussian distributions, exact results can be found. Usually, however, we rely on the χ^2 approximation for differences between deviances for nested models. See appendices A and C. In some circumstances the deviance itself may be approximated by χ^2, for example in discrete data problems where the counts are large. In general, however, the χ^2 approximations for the deviance are not very good even as $n \to \infty$. Further work on the asymptotic distribution of $D(\mathbf{Y}; \hat{\boldsymbol{\mu}})$ remains to be done. The analysis-of-deviance table is best regarded as a screening device for picking out obviously important terms, no attempt being made to assign precise significance levels to the raw deviances.

2.4 Residuals

For Normal models we can express the dependent variate in the form

$$y = \hat{\mu} + (y - \hat{\mu}),$$

i.e. datum = fitted value + residual. Residuals can be used to explore the adequacy of fit of a model, in respect of choice of variance function, link function and terms in the linear predictor. Residuals may also indicate the presence of anomalous values requiring further investigation (see Chapter 12). For generalized linear models we require an extended definition of residuals, applicable to all the distributions that may replace the Normal. It is convenient if these residuals can be used for the same purposes as standard Normal residuals.

In the following section, we use the theoretical form, involving μ rather than $\hat{\mu}$, and we define three forms of generalized residual, which we call the Pearson, Anscombe and deviance residuals.

2.4.1 *Pearson residual*

The Pearson residual, defined by

$$r_{\mathrm{P}} = \frac{y - \mu}{\sqrt{V(\mu)}}, \tag{2.11}$$

is just the raw residual scaled by the estimated standard deviation of Y. The name is taken from the fact that for the Poisson distribution the Pearson residual is just the signed square root of the component of the Pearson X^2 goodness-of-fit statistic, so that

$$\sum r_{\mathrm{P}}^2 = X^2.$$

However Pearson's statistic is used in this book not so much as a goodness-of-fit statistic but as a measure of residual variation.

2.4.2 *Anscombe residual*

A disadvantage of the Pearson residual is that the distribution of r_P for non-Normal distributions is often markedly skewed, and so it may fail to have properties similar to those of a Normal-theory residual. Anscombe proposed defining a residual using a function $A(y)$ in place of y, where $A(\cdot)$ is chosen to make the distribution of $A(Y)$ 'as Normal as possible'. Wedderburn (unpublished, but see Barndorff-Nielsen, 1978) showed that, for the likelihood functions occurring in generalized linear models, the function $A(\cdot)$ is given by

$$A(\cdot) = \int \frac{d\mu}{V^{1/3}(\mu)}.$$

Thus for the Poisson distribution we have

$$\int \frac{d\mu}{\mu^{1/3}} = \tfrac{3}{2}\mu^{2/3},$$

so that we base our residual on $y^{2/3} - \mu^{2/3}$. Now the transformation that 'Normalizes' the probability function does not at the same time stabilize the variance, so that we must scale by dividing by the square root of the variance of $A(Y)$, which is, to the first order, $A'(\mu)\sqrt{V(\mu)}$. Thus for the Poisson distribution the Anscombe residual, to be denoted by r_A, is given by

$$r_A = \frac{\tfrac{3}{2}(y^{2/3} - \mu^{2/3})}{\mu^{1/6}}.$$

See Anscombe (1953) and Cox and Snell (1968) for the definition of the corresponding residual for the binomial distribution. For the gamma distribution the Anscombe residual takes the form

$$r_A = \frac{3(y^{1/3} - \mu^{1/3})}{\mu^{1/3}}.$$

This cube-root transformation was used by Wilson and Hilferty (1931) to normalize variables with a χ^2 distribution. Similarly the inverse Gaussian distribution gives

$$r_A = (\log y - \log \mu)/\mu^{1/2}.$$

Table 2.2 *Comparison of Poisson residuals*

	r_A	r_D	r_P
c	$\frac{3}{2}(c^{2/3}-1)$	$\{2(c\log c - c + 1)\}^{1/2}$	$c-1$
0.0	−1.5	−1.414	−1.0
0.2	−0.987	−0.956	−0.8
0.4	−0.686	−0.683	−0.6
0.6	−0.433	−0.432	−0.2
1.0	0.0	0.0	0.0
1.5	0.466	0.465	0.5
2.0	0.881	0.879	1.0
2.5	1.263	1.258	1.5
3.0	1.620	1.610	2.0
4.0	2.280	2.256	3.0
5.0	2.886	2.845	4.0
10.0	5.462	5.296	9.0

2.4.3 *Deviance residual*

If the deviance is used as a measure of discrepancy of a generalized linear model, then each unit contributes a quantity d_i to that measure, so that $\sum d_i = D$. Hence if we define

$$r_D = \text{sign}(y-\mu)\sqrt{d_i},$$

we have a quantity that increases with $y_i - \mu_i$ and for which $\sum r_D^2 = D$. Thus for the Poisson distribution we have

$$r_D = \text{sign}(y-\mu)\{2(y\log(y/\mu) - y + \mu)\}^{1/2}.$$

Although the Anscombe and deviance residuals appear to have very different functional forms for non-Normal distributions, the values that they take for given y and μ are often remarkably similar, as is clear from a Taylor series expansion. Consider again the Poisson distribution and set $y = c\mu$, so that

$$r_A = \tfrac{3}{2}\mu^{1/2}(c^{2/3}-1)$$

and

$$r_D = \text{sign}(c-1)\mu^{1/2}[2(c\log c - c + 1)]^{1/2}.$$

Table 2.2 shows that the two functions $\frac{3}{2}(c^{2/3}-1)$ and $[2(c\log c - c + 1)]^{1/2}$ are numerically very similar for a range of values of c.

Within this range the maximum difference between r_A and r_D is about 6% at $c = 0$, and much less over most of the range. The Pearson residual is considerably greater in the upper part of the range but goes less far in the negative direction.

For a more extensive examination of definitions of residuals in exponential-family models, see Pierce and Schafer (1986).

2.5 An algorithm for fitting generalized linear models

We shall show that the maximum-likelihood estimates of the parameters $\boldsymbol{\beta}$ in the linear predictor η can be obtained by iterative weighted least squares. In this regression the dependent variable is not y but z, a linearized form of the link function applied to y, and the weights are functions of the fitted values $\hat{\mu}$. The process is iterative because both the *adjusted dependent variable* z and the weight W depend on the fitted values, for which only current estimates are available. The procedure underlying the iteration is as follows. Let $\hat{\boldsymbol{\eta}}_0$ be the current estimate of the linear predictor, with corresponding fitted value $\hat{\boldsymbol{\mu}}_0$ derived from the link function $\eta = g(\mu)$. Form the adjusted dependent variate with typical value

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)_0,$$

where the derivative of the link is evaluated at $\hat{\mu}_0$. The quadratic weight is defined by

$$W_0^{-1} = \left(\frac{d\eta}{d\mu}\right)_0^2 V_0, \qquad (2.12)$$

where V_0 is the variance function evaluated at $\hat{\mu}_0$. Now regress z_0 on the covariates $x_1, \ldots, x_p$ with weight W_0 to give new estimates $\hat{\boldsymbol{\beta}}_1$ of the parameters; from these form a new estimate $\hat{\boldsymbol{\eta}}_1$, of the linear predictor. Repeat until changes are sufficiently small.

Note that z is just a linearized form of the link function applied to the data, for, to first order,

$$g(y) \simeq g(\mu) + (y - \mu)g'(\mu)$$

and the right-hand side is

$$\eta + (y - \mu)\frac{d\eta}{d\mu}.$$

The variance of Z is just W^{-1} (ignoring the dispersion parameter), assuming that η and μ are fixed and known. In this formulation the way in which the calculations for the regression are to be done is left open; we discuss some possibilities in section 3.8.

A convenient feature of this algorithm is that it suggests a simple starting procedure to get the iteration under way. This consists of using the data themselves as the first estimate of $\hat{\mu}_0$ and from this deriving $\hat{\eta}_0$, $(d\eta/d\mu)_0$ and V_0. Adjustments may be required to the data to prevent, for example, our trying to evaluate log(0) as the starting value for η when the log link is applied to counts whose value is zero. These adjustments are described in the appropriate chapters, as will various complexities sometimes associated with the convergence of the iterative process.

2.5.1 *Justification of the fitting procedure*

We first show that the maximum-likelihood equations for β_j are given by

$$\sum W(y-\mu)\,\frac{d\eta}{d\mu}\,x_j = 0 \qquad (2.13)$$

for each covariate x_j, where $\sum$ without a suffix denotes summation over the units, and W is defined in equation (2.12) above.

The log likelihood for a single observation, in canonical form, is given by

$$l = \{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)$$

and we require an expression for $\partial l/\partial\beta_j$. Now, by the chain rule,

$$\frac{\partial l}{\partial\beta_j} = \frac{\partial l}{\partial\theta}\,\frac{d\theta}{d\mu}\,\frac{d\mu}{d\eta}\,\frac{\partial\eta}{\partial\beta_j}.$$

From $b'(\theta) = \mu$ and $b''(\theta) = V$ we derive $d\mu/d\theta = V$, and from $\eta = \sum\beta_j x_j$ we get $\partial\eta/\partial\beta_j = x_j$. Therefore

$$\begin{aligned}\frac{\partial l}{\partial\beta_j} &= \frac{(y-\mu)}{a(\phi)}\,\frac{1}{V}\,\frac{d\mu}{d\eta}\,x_j\\ &= \frac{W}{a(\phi)}(y-\mu)\frac{d\eta}{d\mu}\,x_j\end{aligned}$$

from (2.12).

With constant dispersion ($a(\phi) = \phi$), the factor $a(\phi)$ disappears and we arrive at (2.13) after summing over the observations. With unequal prior weights, giving a dispersion of the form ϕ/w, an extra factor w enters (2.13).

Fisher's scoring method uses the gradient vector $\partial l/\partial \boldsymbol{\beta} = \mathbf{u}$, say, and minus the expected value of the Hessian matrix

$$-E\left(\frac{\partial^2 l}{\partial\beta_r \partial\beta_s}\right) = \mathbf{A}, \quad \text{say.}$$

Given the current estimate $\mathbf{b}$ of $\boldsymbol{\beta}$, we derive an adjustment $\delta\mathbf{b}$ defined as the solution of

$$\mathbf{A}\,\delta\mathbf{b} = \mathbf{u}.$$

Now the components of $\mathbf{u}$ (omitting the dispersion factor) are

$$u_r = \sum W\,(y-\mu)\,\frac{d\eta}{d\mu}\,x_r,$$

so that

$$\begin{aligned} A_{rs} &= -E\frac{\partial u_r}{\partial \beta_s} \\ &= -E\sum\left[(y-\mu)\frac{\partial}{\partial\beta_s}\left\{W\frac{d\eta}{d\mu}x_r\right\} + W\frac{d\eta}{d\mu}x_r\,\frac{\partial}{\partial\beta_s}\,(y-\mu)\right] \quad (2.14) \end{aligned}$$

The first term vanishes on taking expectations while the second reduces to

$$\sum_i W\,\frac{d\eta}{d\mu}\,x_r\,\frac{\partial\mu}{\partial\beta_s} = \sum_i W x_r x_s.$$

Thus $\mathbf{A}$ is the weighted sums-of-squares-and-products matrix of the covariates with weights W.

The new estimate $\mathbf{b}^* = \mathbf{b} + \delta\mathbf{b}$ of $\boldsymbol{\beta}$ thus satisfies the equation

$$\mathbf{A}\mathbf{b}^* = \mathbf{A}\mathbf{b} + \mathbf{A}\,\delta\mathbf{b} = \mathbf{A}\mathbf{b} + \mathbf{u}.$$

Now

$$(\mathbf{A}\mathbf{b})_r = \sum_s A_{rs} b_s = \sum W x_r \eta.$$

Thus the new estimate $\mathbf{b}^*$ satisfies

$$(\mathbf{Ab}^*)_r = \sum_i Wx_r\{\eta + (y - \mu)d\eta/d\mu\},$$

where the sum is over the n units. These equations have the form of linear weighted least-squares equations with weight

$$W = V^{-1}\left(\frac{d\mu}{d\eta}\right)^2$$

and dependent variate

$$z = \eta + (y - \mu)\frac{d\eta}{d\mu}.$$

Note that simplification occurs for the canonical links where the expected value and the actual value of the Hessian matrix coincide, so that the Fisher scoring method and the Newton-Raphson method reduce to the same algorithm. This comes about because the linear weight function $Wd\eta/d\mu$ in the maximum-likelihood equations (2.13) is a constant, so that the first term in the expansion of the Hessian (2.14) is identically zero. Note also that $W = V$ for this case. Finally, if the model is linear on the scale on which Fisher's information is constant, i.e. $g'(\mu) = V^{-\frac{1}{2}}(\mu)$, the vector of weights is constant and need not be recomputed at each iteration.

2.6 Bibliographic notes

The fitting of generalized linear models is accomplished here using a variant of the Newton-Raphson method known as the scoring method. This variation was first introduced in the context of probit analysis by Fisher (1935) in the appendix of a paper by Bliss (1935). Details are given by Finney (1971). For further discussion and extensions see Green (1984) and Jørgensen (1984).

The term 'generalized linear model' is due to Nelder and Wedderburn (1972), who extended the scoring method to deal with maximum-likelihood estimation for exponential-family models. See also Bradley (1973) and Jennrich and Moore (1975).

Linear exponential family models (with canonical link) have been studied by Dempster (1971), Berk (1972), and Haberman (1977). For an extensive rigorous mathematical treatment see Barndorff-Nielsen (1978). Important special cases of linear exponential family models have been considered by Cox (1970) and by Breslow (1976).

2.7 Further results and exercises 2

2.1 Let $f_0(y)$ be an arbitrary density or probability distribution having moment generating function

$$M(\xi) = E\{\exp(\xi Y)\} = \exp\{b(\xi)\},$$

assumed finite for a range of ξ-values that includes 0. Now consider the exponentially weighted density

$$f_Y(y;\theta) \propto \exp(\theta y) f_0(y).$$

Derive the normalization factor for the weighted density and show that $f_Y(y;\theta)$ has the exponential-family form (2.4) with $a(\phi) = 1$.

2.2 Show that the cumulants of the weighted density $f_Y(y;\theta)$ are given by

$$\bar{\kappa}_r = b^{(r)}(\theta),$$

whereas the cumulants of the initial density are $b^{(r)}(0)$.

2.3 Let $Y_1, \ldots, Y_\nu$ be ν independent copies of the random variable Y having the weighted density function $f_Y(y;\theta)$. Show that the arithmetic mean $\bar{Y} = (Y_1 + \ldots + Y_\nu)/\nu$ has a density of the form (2.4) with $a(\phi) = \nu^{-1}$. Show also that the cumulants of $\bar{Y}$ are

$$\kappa_r(\bar{Y}) = b^{(r)}(\theta)/\nu^{r-1}.$$

Hence establish a central-limit theorem for densities of the form (2.4). [Jørgensen, 1987].

2.4 Discuss the limitations of the averaging operation as a way of generating a two-parameter family of distributions suitable for statistical work. Consider in particular the following points:

1. parameter interpretation,
2. possible non-integer values of ν,
3. dependence of the support of $\bar{Y}$ or $\nu\bar{Y}$ on the parameters, particularly where $f_0(y)$ is discrete.

2.5 Go through the calculations indicated in the previous four exercises, beginning with the distribution $f_0(y)$, which attaches probability one half to $y = 0$ and $y = 1$. What is the distribution of $\nu\bar{Y}$?

2.6 Beginning with the discrete distribution $f_0(y) \propto 1/y!$ for $y = 1, 2, \ldots$, derive the corresponding exponential family by going through the calculations of Exercises 2.1–2.3. Find the cumulant function $b(\theta)$ and hence derive the likelihood equation for $\hat{\theta}$ based on a sample of independent and identically distributed observations.

2.7 For the distribution (2.4), show that the rth cumulant of Y is

$$\kappa_r = b^{(r)}(\theta) \times a^{r-1}(\phi).$$

Hence deduce that

$$\kappa_3 = \kappa_2 \kappa_2' \quad \text{and} \quad \kappa_4 = \kappa_2 \kappa_3',$$

where primes denote differentiation with respect to μ.

2.8 Show that

$$f_X(x; \theta, \nu) = \frac{(1 - x^2)^{\nu - 1/2}}{(1 - 2\theta x + \theta^2)^{\nu} B(\nu + \frac{1}{2}, \frac{1}{2})} \qquad -1 \leq x \leq 1,$$

is a probability density on $(-1, 1)$ for all parameter values $\nu > -\frac{1}{2}$, $-1 \leq \theta \leq 1$ (McCullagh, 1989). [If all efforts at integration fail, check that the claim is true for $\theta = \pm 1, 0$ and, by numerical integration using Simpson's rule or other Newton-Cotes formula, for other values of (θ, ν).]

Sketch the density for $\theta = 0, \pm\frac{1}{2}, \pm 1$, $\nu = 3$.

2.9 For the density given above, show that for all $r > -\nu$,

$$E\left(\frac{1 - X^2}{1 - 2\theta X + \theta^2}\right)^r = \frac{B(\nu + r + \frac{1}{2}, \frac{1}{2})}{B(\nu + \frac{1}{2}, \frac{1}{2})},$$

$$E\left(\frac{X - \theta}{1 - 2\theta X + \theta^2}\right) = 0.$$

Hence deduce that $T(\theta) = (1 - X^2)/(1 - 2\theta X + \theta^2)$ is a pivotal statistic whose distribution does not depend on θ. Find the distribution of T.

2.10 Show that for fixed θ the density $f_X(x; \theta, \nu)$ given above is of the exponential-family type (2.4) with $\phi = 1$, $y = \log T(\theta)$ and canonical parameter ν. Find the cumulant function $b(\cdot)$.

2.11 Show that $-2\nu \log T(\theta_0)$ is the scaled deviance statistic for testing the hypothesis $H_0 : \theta = \theta_0$ on the basis of a single observation X. Deduce that for large ν and under H_0

$$-(2\nu + \tfrac{1}{2}) \log T(\theta_0) \sim \chi_1^2$$

approximately.

2.12 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed with density $f_X(x; \theta, \nu)$ as given above. Show that $\hat{\theta}_\nu$, the maximum-likelihood estimate of θ for fixed ν, is independent of ν. Calculate the Fisher information for (θ, ν) and show that this matrix is diagonal.

2.13 Using the result given in Exercise 2.8 show that

$$f_X(x; \theta, \nu) = \frac{(1 - x^2)^{\nu - 1/2}\, |\theta|^{2\nu}}{(1 - 2\theta x + \theta^2)^\nu\, B(\nu + \frac{1}{2}, \frac{1}{2})} \qquad \nu > -\tfrac{1}{2},\ |\theta| \geq 1,$$

is a probability density on the interval $-1 \leq x \leq 1$ for the parameter values indicated. Comment on the behaviour of the likelihood function and the Fisher information near $\theta = \pm 1$. [McCullagh, 1989].

2.14 In order to construct a family of the type (2.4), suppose we begin with the logistic density

$$f_0(x) = \frac{e^x}{(1 + e^x)^2} = \frac{1}{\left(2\cosh(x/2)\right)^2} \quad \text{for} \quad -\infty < x < \infty.$$

Show that the associated exponential family, also known as the exponentially tilted family, is

$$f(x; \theta) = \frac{e^{x(1+\theta)}}{(1 + e^x)^2\, \Gamma(1 + \theta)\Gamma(1 - \theta)} = \frac{e^{x(1+\theta)} \sin(\pi\theta)}{(1 + e^x)^2\, \pi\theta}$$

for $-1 < \theta < 1$. Deduce that $f(x;\theta) = f(-x;-\theta)$. Plot the density for $\theta = 0.25$, 0.5 and 0.75. Find the cumulant function $b(\theta)$ and show that the mean of the tilted density is

$$E(X;\theta) = b'(\theta) = \frac{1}{\theta} - \pi\cot(\pi\theta).$$

Plot $E(X;\theta)$ against θ to show that the mean-value parameter is a monotone function of the canonical parameter.

For what values of θ does $\exp(X)$ have an F-distribution?

2.15 Discuss the connection between the above exponential family and the family generated by the particular hyperbolic secant density

$$f_2(x;0) = \frac{x}{2\sinh(\pi x/2)} \quad \text{for} \quad -\infty < x < \infty,$$

whose cumulant generating function is $-2\log\cos\theta$ for $|\theta| < \pi/2$.

Find the mean and variance of the tilted density as functions of θ. Plot the exponentially tilted density $f_2(x;\theta)$ for $\theta = 0$, $\pi/6$ and $\pi/3$. [Morris 1982, section 5.]

2.16 Suppose that $b_0(\theta)$, the cumulant generating function for the density $f_0(x)$, is defined for $-1 < \theta < 1$. Show that the cumulant generating function of $f(x;\mu,\sigma) = f_0((x-\mu)/\sigma)/\sigma$ is given by

$$b(\theta;\mu,\sigma) = \theta\mu + b_0(\sigma\theta)$$

for $-1 < \sigma\theta < 1$. Hence deduce that location-scale transformation and exponential tilting are commutative operations but that the parameters θ, σ are not variation independent.

For the special case $b_0(\theta) = \frac{1}{2}\theta^2$, show that the three parameters (μ, σ, θ) are not all identifiable.