J. R. Statist. Soc. A,
(1972), **135,** *Part* 3, p. 370

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

Rothamsted Experimental Station, Harpenden, Herts

SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

Keywords: ANALYSIS OF VARIANCE; CONTINGENCY TABLES; EXPONENTIAL FAMILIES; INVERSE POLYNOMIALS; LINEAR MODELS; MAXIMUM LIKELIHOOD; QUANTAL RESPONSE; REGRESSION; VARIANCE COMPONENTS; WEIGHTED LEAST SQUARES

INTRODUCTION

LINEAR models customarily embody both systematic and random (error) components, with the errors usually assumed to have normal distributions. The associated analytic technique is least-squares theory, which in its classical form assumed just one error component; extensions for multiple errors have been developed primarily for analysis of designed experiments and survey data. Techniques developed for non-normal data include probit analysis, where a binomial variate has a parameter related to an assumed underlying tolerance distribution, and contingency tables, where the distribution is multinomial and the systematic part of the model usually multiplicative. In both these examples there is a linear aspect to the model; thus in probit analysis the parameter p is a function of tolerance Y which is itself linear on the dose (or some function thereof), and in a contingency table with a multiplicative model the logarithm of the expected probability is assumed linear on classifying factors defining the table. Thus for both, the systematic part of the model has a linear basis. In another extension (Nelder, 1968) a certain transformation is used to produce normal errors, and a different transformation of the expected values is used to produce linearity.

So far we have mentioned models associated with the normal, binomial and multinomial distributions (this last can be thought of as a set of Poisson distributions with constraints). A further class is based on the χ^2 or gamma distribution and arises in the estimation of variance components from independent quadratic forms derived from the original observations. Again the systematic component of the model has a linear structure.

In this paper we develop a class of *generalized linear models*, which includes all the above examples, and we give a unified procedure for fitting them based on likelihood. This procedure is a generalization of the well-known one described by Finney (1952) for maximum likelihood estimation in probit analysis. Section 1 defines the models, and Section 2 develops the fitting process and generalizes the analysis of variance. Section 3 gives examples with four special distributions for the random components. In Section 4 we consider the usefulness of the models for courses of instruction in statistics.

1.1. *The Random Component*

Suppose our observations z come from a distribution with density function

$$\pi(z;\ \theta,\phi) = \exp\left[\alpha(\phi)\{z\theta - g(\theta) + h(z)\} + \beta(\phi, z)\right],$$

where $\alpha(\phi) > 0$ so that for fixed ϕ we have an exponential family. The parameter ϕ could stand for a certain type of nuisance parameter such as the variance σ^2 of a normal distribution or the parameter p of a gamma distribution (see Section 3.4). We denote the mean of z by μ.

We require expressions for the first and second derivatives of the log-likelihood in terms of the mean and variance of z and the scale factor $\alpha(\phi)$. We use the results (see, for example, Kendall and Stuart, 1967, p. 9)

$$E(\partial L/\partial\theta) = 0 \tag{1}$$

and

$$E(\partial^2 L/\partial\theta^2) = -E(\partial L/\partial\theta)^2, \tag{2}$$

where the differentiation under the sign of integration used in their derivation can be justified by Theorem 9 of Lehmann (1959).

We have

$$\partial L/\partial\theta = \alpha(\phi)\{z - g'(\theta)\}.$$

Then (1) implies that

$$\mu = E(z) = g'(\theta);$$

hence

$$\partial L/\partial\theta = \alpha(\phi)(z-\mu). \tag{3}$$

From (2) we obtain $\alpha(\phi)g''(\theta) = [\alpha(\phi)]^2 \operatorname{var}(z)$, whence

$$g''(\theta) = \alpha(\phi)\operatorname{var}(z) = V, \quad \text{say}, \tag{4}$$

so that V is the variance of z when the scale factor is unity. Then

$$\partial^2 L/\partial\theta^2 = -\alpha(\phi)V. \tag{5}$$

We note also that

$$V = d\mu/d\theta. \tag{6}$$

For a one-parameter exponential family $\alpha(\phi) = 1$, we can write

$$\pi(z;\ \theta) = \exp\{z\theta - g(\theta) + h(z)\},$$

so that

$$\partial L/\partial\theta = z - \mu$$

and

$$-\partial^2 L/\partial\theta^2 = V = \operatorname{var}(z).$$

1.2. *The Linear Model for Systematic Effects*

The term "linear model" usually encompasses both systematic and random components in a statistical model, but we shall restrict the term to include only the systematic components. We write

$$Y = \sum_{i=1}^{m} \beta_i x_i$$

when the x_i are independent variates whose values are supposed known and β_i are parameters. The β_i may have fixed (known) values or be unknown and require estimation. An independent variate may be *quantitative* and produce a single x-variate in the model, or *qualitative* and produce a set of x-variates whose values are 0 and 1, or *mixed*. Consider the model

$$Y_{ij} = \alpha_i + \beta u_{ij} + \gamma_j v_{ij} \quad (i = 1, \ldots, n, j = 1, \ldots, p),$$

where the data are indexed by factors whose levels are denoted by i and j. The term α_i includes n parameters associated with a qualitative variate represented by n dummy x-variate components taking values 1 for one level and 0 for the rest; βu_{ij} represents a quantitative variate, namely u with single parameter β, and $\gamma_j v_{ij}$ shows p parameters γ_j associated with a mixed independent variate whose p components take the values of v_{ij} for one level of j and zero for the rest. A notation suitable for computer use has been developed by C. E. Rogers and G. N. Wilkinson and is to be published in *Applied Statistics.*

1.3. *The Generalized Linear Model*

We now combine the systematic and random components in our model to produce the generalized linear model. This is characterized by

(i) A dependent variable z whose distribution with parameter θ is one of the class in Section 1.1.

(ii) A set of independent variables $x_1, \ldots, x_m$ and predicted $Y = \sum \beta_i x_i$ as in Section 1.2.

(iii) A linking function $\theta = f(Y)$ connecting the parameter θ of the distribution of z with the Y's of the linear model.

When z is normally distributed with mean θ and variance σ^2 and when $\theta = Y$, we have ordinary linear models with Normal errors. Other examples of these models will be described in Section 3 under the various distributions of the exponential type. We now consider the solution of the maximum likelihood equations for the parameters of the generalized linear models and show its equivalence to a procedure of iterative weighted least squares.

2. FITTING THE MODELS

2.1. *The Maximum Likelihood Equations*

The solution of the maximum likelihood equations is equivalent to an iterative weighted least-squares procedure with a weight function

$$w = (d\mu/dY)^2/V$$

and a modified dependendent variable (the working probit of probit analysis)

$$y = Y + (z-\mu)/(d\mu/dY),$$

where μ, Y and V are based on current estimates. This generalizes the results of Nelder (1968).

Proof. Writing L for the log-likelihood from one observation we have, from (3) and (6),

$$\frac{\partial L}{\partial \beta_i} = \alpha(\theta)(z-\mu)\frac{d\theta}{d\mu}\frac{d\mu}{dY}x_i$$

$$= \alpha(\theta)\frac{z-\mu}{V}\frac{d\mu}{dY}x_i \tag{7}$$

and

$$\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} = \frac{\partial^2 L}{\partial Y^2}x_i x_j,$$

where

$$\frac{\partial^2 Z}{\partial Y^2} = \frac{\partial^2 L}{\partial \theta^2}\left(\frac{d\theta}{dY}\right)^2 + \frac{\partial L}{\partial \theta}\frac{d^2\theta}{dY^2}$$

$$= \alpha(\phi)\left\{-V\left(\frac{d\theta}{d\mu}\right)^2\left(\frac{d\mu}{dY}\right)^2 + (z-\mu)\frac{d^2\theta}{dY^2}\right\}$$

$$= \alpha(\phi)\left\{-\left(\frac{d\mu}{dY}\right)^2\Big/ V + (z-\mu)\frac{d^2\theta}{dY^2}\right\}. \tag{8}$$

The expected second derivative with negative sign is thus

$$\alpha(\phi)\left\{\left(\frac{d\mu}{dY}\right)^2\Big/ V\right\}x_i x_j \quad \text{from (8).}$$

Writing w for the weight function $(d\mu/dy)^2/V$, (7) gives

$$\partial L/\partial\beta_i = \alpha(\phi)\, w x_i(z-\mu)/(d\mu/dY).$$

Thus the Newton–Raphson process with expected second derivatives (equivalent to Fisher's scoring technique) for a sample of n gives

$$\mathbf{A}\delta\boldsymbol{\beta} = \mathbf{C}, \tag{9}$$

where $\mathbf{A}$ is a $m \times m$ matrix with

$$A_{ij} = \sum_{k=1}^{n} w_k x_{ik} x_{jk}$$

and $\mathbf{C}$ is a $m \times 1$ vector with

$$C_i = \sum w_k x_{ik}(z-\mu)/(d\mu/dY).$$

Finally we have

$$(A\beta)_i = \sum A_{ij}\beta_j = \sum w_k x_{ik} Y_k$$

so that (9) may be written in the form

$$\mathbf{A}\boldsymbol{\beta}^* = \mathbf{r}$$

where $r_i = \sum w_k x_{ik} y_k$, $y_k = Y_k + (z_k - \mu_k)/(d\mu_k/dY_k)$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta} + \delta\boldsymbol{\beta}$.

Starting method

In practice we can obtain a good starting procedure for iteration as follows: take as a first approximation $\mu = z$ and calculate Y from it; then calculate w as before and set $y = Y$. Then obtain the first approximation to the β's by regression. The method may need slight modification to deal with extreme values of z. For instance, with the binomial distribution it will probably be adequate to replace instances of $z = 0$ or $z = n$ with $z = \frac{1}{2}$ and $z = n - \frac{1}{2}$, where, e.g. with the probit and logit transformations, $\mu = 0$ or $\mu = n$ would lead to infinite values for Y.

2.2. *Sufficient Statistics*

An important special case occurs when θ, the parameter of the distribution of the random element, and Y the predicted value of the linear model, coincide. Then

$$L = zY - g(Y) + h(z),$$

and, using (3),

$$\partial L / \partial \beta_i = \alpha(\phi)(z - \mu) x_i.$$

The maximum likelihood equations are then of the form $\sum_k (z - \hat{\mu}) x_{ik} = 0$, the summation being over the observations. Hence we have

$$\sum_k z_k x_{ik} = \sum_k \hat{\mu}_k x_{ik}. \tag{10}$$

For a qualitative independent variate, this implies that the fitted marginal totals with respect to that variate will be equal to the observed ones.

From the expression for L we see that the quantities $\sum_k z_k x_{ik}$ are a set of sufficient statistics. Also, in (8) $d^2\theta/dY^2 = 0$ and so

$$\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} = E\left(\frac{\partial^2 L}{\partial \beta_i \partial \beta_j}\right) = -\alpha(\phi)\left(\frac{d\mu^2}{dy}\bigg/ V\right) x_i x_j. \tag{11}$$

When θ is also the mean of the distribution, i.e. $\mu = \theta = Y$, we have the usual linear model with normal errors, for $g'(\theta) = \theta$ gives

$$g(\theta) = \tfrac{1}{2}\theta^2 + \text{const}$$

which uniquely determines the distribution as Normal with variance $1/\alpha(\phi)$ (using Theorem 1 of Patil and Shorrock, 1965). The sub-class of models for which there are sufficient statistics was noted by Cox (1968), and Dempster (1971) has extended it to include many dependent variates.

2.3. *The Analysis of Deviance*

A linear model is said to be *ordered* if the fitting of the β's is to be done in the same sequence as their declaration in the model. Ordering (or partial ordering) may be implied by the structure of the model; for instance it makes no sense to fit an interaction term $(ab)_{ij}$ before fitting the corresponding main effects a_i and b_j. It may also be implied by the objectives of the fitting, i.e. if a trend must be removed first before the fitting of further effects. More commonly, however, the ordering is to some extent arbitrary, and this gives rise to difficult problems of inference which we shall not try to tackle here. For ease of exposition of the basic ideas we shall assume that the model under consideration is ordered, and will be fitted sequentially a term at a time. The objectives of the fitting will be to assess how many terms are required for an adequate description of the data, and to derive the associated estimates of the parameters and their information matrix.

Two extreme models are conceivable for any set of data, the *minimal model* which contains the smallest set of terms that the problem allows, and the *complete model* in which all the Ys are different and match the data completely so that $\hat{\mu} = z$. An extreme case of the minimal model is the null model, which is equivalent to fitting the grand mean only and effectively consigns all the variation in the data to the random component of the model, while the complete model fits exactly and so consigns all the variation in the data to the systematic part. The model-fitting process with an ordered model thus consists of proceeding a suitable distance from the minimal model towards the complete model. At each stage we trade increasing goodness-of-fit to the current set of data against increasing complexity of the model. The fitting of the parameters at each stage is done by maximizing the likelihood for the current model and the matching of the model to the data will be measured quantitatively by the quantity $-2L_{\max}$ which we propose to call the *deviance*. For the four special distributions the deviance takes the form:

$$\begin{array}{ll} \text{Normal} & \sum(z-\hat{\mu})^2/\sigma^2, \\ \text{Poisson} & 2\{\sum z\ln(z/\hat{\mu})-\sum(z-\hat{\mu})\}, \\ \text{Binomial} & 2[\sum z\ln(z/\hat{\mu})+\sum(n-z)\ln\{(n-z)/(n-\hat{\mu})\}], \\ \text{Gamma} & 2p\{-\sum\ln(z/\hat{\mu})+\sum(z-\hat{\mu})/\hat{\mu}\}. \end{array}$$

Note that the deviance is measured from that of the complete model, so that terms involving constants, the data alone, or the scale factor alone are omitted. The second term in the expressions for the Poisson and gamma distribution is commonly identically zero (see Appendix for conditions and proof).

Associated with each model is a quantity r termed *the degrees of freedom* which is given by the rank of the X matrix, or equivalently the number of linearly independent parameters to be estimated. For a sample of n independent observations, the deviance for the model has residual degrees of freedom $(n-r)$. The degrees of freedom, multiplied where necessary by a scale factor, form a scale for a set of sequential models with which deviances can be compared; when (residual degrees of freedom × scale factor) is approximately equal to the deviance of the current model then it is unlikely that further fitting of systematic components is worth while. The scale factor may be known (e.g. unity for the Poisson distribution) or unknown (e.g. for the normal distribution with unknown variance) If unknown it may be estimable directly, e.g. by replicate observations, or indirectly from the deviance after an adequate model has been fitted. The adequacy of the model may be determined by plotting successive deviances against their degrees of freedom, and accepting as a measure of the scale factor the linear portion through the origin determined by those points with fewest degrees of freedom.

2.4. *The Generalization of Analysis of Variance*

The first differences of the deviances for the normal distribution are (apart from a scale factor) the sums of squares in the analysis of variance for a sequential fit as shown for a three-term model in Table 1.

TABLE 1

Deviances and their differences

Model term	*Deviance*	*Difference*	*Component*
Minimal	d_m		
		$d_m - d_A$	A
A	d_A		
		$d_A - d_{AB}$	B eliminating A
B	d_{AB}		
		$d_{AB} - d_{ABC}$	C eliminating A and B
C	d_{ABC}		
		$d_{ABC} - d_0$	Residual
Complete	d_0		

The generalized analysis of variance for a sequential model is now defined to have components given by the first differences of the deviance, with degrees of freedom defined as above. These components have distributions proportional to χ^2, exactly for normal errors, approximately for others. Such a generalization of the analysis of variance was suggested by Good (1967).

3. SPECIAL DISTRIBUTIONS

3.1. *The Normal Distribution*

Here, we have

$$L = \sigma^{-2}(z\mu - \tfrac{1}{2}\mu^2 - \tfrac{1}{2}z^2) - \tfrac{1}{2}\ln\sigma^2$$

and in the notation of Section 1.2

$$\mu = \theta, \quad V = 1.$$

Inverse polynomials provide an example where we assume that the observations u are normal on the log scale and the systematic effects additive on the inverse scale. Then

$$z = \ln u \quad \text{and} \quad Y = e^{\mu}.$$

Nelder (1966) gives examples of inverse polynomials calculated using the first approximation of the method in this paper.

More generally, as shown in Nelder (1968), we can consider models in which there is a linearizing transformation f and a normalizing transformation g. This means that if the observations are denoted by u, then $g(u)$ is normally distributed with mean μ and constant variance σ^2 and $f\{g^{-1}(\mu)\} = \sum\beta_i x_i$.

Then we have $V = 1$ and $Y = f\{g^{-1}(\mu)\}$ so that

$$w = [(d/dY)\, g\{f^{-1}(Y)\}]^2$$

and

$$y = Y + \{g(u) - \mu\}/[(d/dY)\, g\{f^{-1}(Y)\}].$$

Example: Fisher's tuberculin-test data

Fisher (1949) published 16 measurements of tuberculin response which were classified by three four-level factors in a Latin-square type of arrangement as in Table 2.

TABLE 2

Sites	*Cow class*				*Treatments*			
	I	*III*	*II*	*IV*				
3+6	454	249	349	249	A	B	C	D
4+5	408	322	312	347	B	A	D	C
1+8	523	268	411	285	C	D	A	B
2+7	364	283	266	290	D	C	B	A

Fisher gave reasons for believing that the variances of the observations were proportional to their expectation, and that the systematic part of the model was linear on the log-scale. The treatments were:

B	Standard single
A	Standard double
D	Weybridge half
C	Weybridge single

and these were treated as a 2×2 factorial arrangement, no interaction being fitted.

If the data had been Poisson observations the maximum likelihood estimates would have had the property that the marginal totals of the fitted values (on the untransformed scale) would be equal to the marginal totals of the observations. Although the observations were not, in fact, counts but measurements in millimetres, Fisher decided to estimate the effects as if they were Poisson observations. He produced approximations to the effects by a method which made use of features of the particular Latin square, and then verified that these gave fitted values with margins approximately equal to the observed ones.

Another approach would be to treat the square roots of the observations as normally distributed with variance $\sigma^2/4$. Then we have $z = \surd u$ where u is an observation and

$$Y = 2\ln\mu.$$

Fisher gave estimates of effects on the log scale relative to **B**. We produced estimates by the square-root/logarithmic method just described, and the two sets of estimates are given in Table 3.

TABLE 3

	Fisher's result	*Our result*
B	0·0000	0·0000
A	0·2089	0·2092
D	0·0019	0·0023
C	0·2108	0·2115

The method of this paper can also be used to analyse the data as if they were Poisson observations. The estimates of effects obtained by this method agree with our other estimates to about four decimal places.

3.2. *The Poisson Distribution*

Here $L = z\ln\mu - \mu$ so we have $\theta = \ln\mu$ and $V = \mu$.

When $Y = \ln\mu$ there are sufficient statistics and a unique maximum likelihood estimate of β, provided it is finite. It will always be finite if there are no zero observations.

If $Y = \mu^{\lambda}$ $(0 < \lambda < 1)$, $L \rightarrow -\infty$ as $|\boldsymbol{\beta}| \rightarrow \infty$ and hence L must have a maximum for finite $\boldsymbol{\beta}$. Also

$$\frac{\partial^2 L}{\partial\beta_i\,\partial\beta_j} = \sum \frac{\partial^2 L}{\partial Y^2} x_i x_j.$$

It is easily verified that $\partial^2 L/\partial Y^2 < 0$ and hence L is negative definite. It follows that $\hat{\boldsymbol{\beta}}$ is uniquely determined. When $Y = \mu$ the same result holds provided that the x's are linearly independent when units with $z = 0$ are excluded.

The main application of generalized linear models with Poisson errors is to contingency tables. These arise from data on counts classified by two or more factors, and the literature on them is enormous (see, for example, Simpson, 1951; Ireland and Kullback, 1968; Chapter 8 of Kullback 1968; Ku *et al.*, 1971).

Probabilistic models for contingency tables are built on assumptions of a multinomial distribution or a set of multinomial distributions. As Birch (1963) has shown, the estimation of a set of independent multinomial distributions is equivalent to the estimation of a set of independent Poisson distributions, and in what follows we shall regard a contingency table as a set of independent Poisson distributions.

The systematic part of models of contingency tables is usually multiplicative, and thus gives sufficient statistics with Poisson errors. The model terms usually correspond to qualitative x's, the equivalent of constant fitting, but quantitative terms occur naturally when the classifying factors have an underlying quantitative basis.

Example: A contingency table

Maxwell (1961, pp. 70–72) discusses the analysis of a 5×4 contingency table giving the number of boys with four different ratings for disturbed dreams in five different age groups. The data are given in Table 4. The higher the rating the more the boy suffers from disturbed dreams.

TABLE 4

Age in years	*Rating*				
	4	3	2	1	*Total*
5– 7	7	3	4	7	21
8– 9	13	11	15	10	49
10–11	7	11	9	23	50
12–13	10	12	9	28	59
14–15	3	4	5	32	44
Total	40	41	42	100	223

Here we can fit main effects and a linear × linear interaction using an x-variate of the form uv where $u = -2, -1, 0, 1$ or 2 according to the age group and v is the rating for disturbed dreams.

The estimated linear × linear interaction is −0·205.

We can form an analysis of deviance thus (regarding main effects as fixed):

Model term(s)	*Deviance*	*Difference*	*Degrees of freedom*
Minimal (i.e. main effects only)	32·46	18·38	1
Linear × linear	14·08	14·08	11
Complete	0		

Treating 18·38 and 14·08 as χ^2 variates with 1 and 11 degrees of freedom respectively we find that 18·38 is large while 14·08 is close to expectation. We conclude that the data are adequately described by a negative linear × linear interaction (indicating that the dream rating tends to decrease with age).

Maxwell, using the method of Yates (1948), obtained a decomposition of a Pearson χ^2 as follows:

Source of variation	*Degrees of freedom*	χ^2
Due to linear regression	1	17·94
Due to departure from linear regression	11	13·73
Total	12	31·67

Maxwell's values of χ^2 are clearly quite close to ours and his conclusions are essentially the same.

3.3. *The Binomial Distribution*

We re-write the usual form

$$L = r \ln p + (n-r) \ln q$$

as

$$L = z \ln(\mu/n) + (n-z) \ln\{(-(\mu/n)\}$$
$$= z \ln\{\mu/(n-\mu)\} + n \ln(n-\mu) + \text{terms in } n,$$

i.e. we put z for r, and $\mu = E(z) = np$. Thus

$$\theta = \ln\{\mu/(n-\mu)\}$$

and

$$V = \mu(n-\mu)/n.$$

Sufficient statistics are provided by the logit transformation giving

$$\mu = ne^y/(1+e^y).$$

Probit analysis

We put

$$\mu = n\Phi(Y),$$

where Φ is the cumulative normal distribution function. There are no sufficient statistics here; the analysis via iterative weighted least-squares is well known (Finney, 1952).

Fitting constants on a logit scale

This technique was introduced by Dyke and Patterson (1952) and is applied to multiway tables of proportions. It is a special case of the logistic transformation when all the x's are qualitative, and yields as sufficient statistics the total responding ($\sum z$) for each relevant margin.

The logit analogue of probit analysis is, of course, formally identical, with quantitative rather than qualitative x's. Again arbitrary mixtures of x types do not introduce anything new. Models based on the logistic transform have been extensively developed by Cox (1970).

From the results of Birch (1963) mentioned above, it follows that models with independent binomial data are equivalent to models with independent Poisson data. Bishop (1969) showed that a binomial model that is additive on the logit scale can be treated as a Poisson model additive on the log scale.

3.4. *The Gamma Distribution*

For the gamma distribution

$$L = -p(z/\mu + \ln \mu - \ln z) - \ln z.$$

We have $\theta = -1/\mu$ and $V = d\mu/d\theta = \mu^2$. There are sufficient statistics when $Y = 1/\mu$.

Thus the inverse transformation to linearity is related to the gamma distribution, as the logarithmic to the Poisson, or the identity transformation to the normal. The corresponding models seem not to have been explored.

One application of linear models involving the gamma distribution is the estimation of variance components. Here we have sums of squares which are proportional to χ^2 variates and the expectation of each is a linear combination of several variances which are to be estimated. It is better to write

$$L = -(z\nu/2\mu) - (\nu/2)\ln \mu + \{(\nu/2)+1\}\ln z,$$

where ν is the degrees of freedom of z; then putting $\theta = -\nu/2\mu$ we have $V = d\mu/d\theta = 2\mu^2/\nu$ and $y = z$, $Y = \mu$ and $w = \nu/2\mu^2$.

The deviance takes the form

$$\sum \nu[-\ln(z/\mu) + \{(z-\hat{\mu})/\hat{\mu}\}]$$

and the result of the Appendix gives $\sum\{\nu(z-\hat{\mu})/\hat{\mu}\} = 0$; hence the deviance simplifies to

$$\sum \nu \ln(\hat{\mu}/z).$$

Example

In a balanced incomplete block design in which $b > v$, we can produce an analysis of variance as follows (Yates, 1940 or paper VIII of Yates, 1970).

	Degrees of freedom
Blocks (eliminating varieties):	
Varietal component	$v-1$
Remainder	$b-v$
Total	$b-1$
Varieties (ignoring blocks)	$v-1$
Intra-block error	$rv-v-b+1$
Total	$rv-1$

The expectations of some of the mean squares are as follows:

	Expected mean square
Blocks (eliminating varieties):	
Varietal component	$\sigma^2 + Ek\sigma_b^2$
Remainder	$\sigma^2 + k\sigma_b^2$
Intra-block error	σ^2

Here σ^2 is the intra-block variance and σ_b^2 is the inter-block variance.

On p. 322 of Yates (1940) (or p. 207 of Yates, 1970) an example is given with $v = 9$, $r = 8$, $k = 4$, $b = 18$, $\lambda = 3$ and $E = \frac{27}{32}$. The three mean squares mentioned above, their degrees of freedom and their expectations are as follows:

	Mean square	*Degrees of freedom*	*Expectation*
Blocks (eliminating varieties):			
Varietal component	4·6329	8	$\sigma^2 + \frac{27}{8}\sigma_b^2$
Remainder	15·3557	9	$\sigma^2 + 4\sigma_b^2$
Intra-block error	2·5968	46	σ^2

Each iteration then takes the form of a weighted fitting of a straight line with x-values $\frac{27}{8}$, 4 and 0.

The estimates obtained were:

$$\hat{\sigma}^2 = 2{\cdot}5870, \quad \hat{\sigma}_b^2 = 2{\cdot}0314.$$

Yates equated the intra-block error mean square and the blocks (eliminating varieties) mean square to their expectations. This gives

$$\hat{\sigma}^2 = 2{\cdot}5968, \quad \hat{\sigma}_b^2 = 2{\cdot}0813.$$

This was one example where the first approximation was not very good. Our first approximation was

$$\hat{\sigma}^2 = 2{\cdot}5874, \quad \hat{\sigma}_b^2 = 0{\cdot}9313$$

Subsequent iterations gave the following values:

Iteration No.	$\hat{\sigma}^2$	$\hat{\sigma}_b^2$
1	2·5695	2·0737
2	2·5874	2·0302
3	2·5870	2·0315
4	2·5870	2·0314

4. THE MODELS IN THE TEACHING OF STATISTICS

We believe that the generalized linear models here developed could form a useful basis for courses in statistics. They give a consistent way of linking together the systematic elements in a model with the random elements. Too often the student meets complex systematic linear models only in connection with normal errors, and if he encounters probit analysis this may seem to have little to do with the linear regression theory he has learnt. By isolating the systematic linear component the student can be introduced to multiway tables and their margins, additivity, weighting, quantitative and qualitative independent variates, and transformations, quite independently of the added complications of errors and associated probability distributions. The essential unity of the linear model, encompassing qualitative, quantitative and mixed independent variates can be brought out and the introduction of qualitative variates brings in naturally the ideas of singularity of matrices and of constraints.

The complementary set of probability distributions would be introduced in the usual way, including the use of transformations of data to attain desirable properties of the errors. The difficult problem of discussing how far transformations can produce both linearity and normality simultaneously now disappears because the models allow two different transformations to be used, one to induce the linearity of the systematic component and one to induce the desired distribution in the error component. (Note that this distribution need not necessarily be equal-variance normal.)

The systematic use of log-likelihood-ratios (or, equivalently, differences in deviance) extends the ideas of analysis of variance to other distributions and produces an additive decomposition for the sequential fit of the model. To appreciate the simplicity that this can produce it is only necessary to look at the algebraic complexities arising from the attempts to analyse contingency tables by extensions of the Pearson χ^2 approach. Irwin (1949), Lancaster (1949, 1950) and Kimball (1954) have given modifications of the usual formulae for the Pearson χ^2 in contingency tables, to produce an additive decomposition of χ^2 when the table is partitioned. These formulae are much more complicated than the usual one for Pearson χ^2. However, the equivalent likelihood statistic, namely

$$2\left\{\sum_{ij} n_{ij} \ln n_{ij} - \sum_{i} n_i \ln n_{i.} - \sum_{i} n_{.j} \ln n_{.j} + n_{..} \ln n_{..}\right\}$$

does not need any modification to make it additive and is easy to compute as tables of $n \ln n$ are available (Kullback, 1968).

Finally, the fact that a single algorithm can be used to fit any of the models implies that quite a small set of routines can provide the basic computing facility to allow students to fit models to a wide range of data. They can get experience of programming by writing special routines to deal with special forms of output required by particular models, such as the LD50 of probit analysis. In this way the distinction can be made between the model-fitting part of an analysis and the subsequent derived quantities (with estimates of their uncertainties) which particular problems require.

We hope that the approach developed in this paper will prove to be a useful way of unifying what are often presented as unrelated statistical procedures, and that this unification will simplify the teaching of the subject to both specialists and non-specialists.

REFERENCES

BIRCH, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc.* B, **25**, 220–233.

BISHOP, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics*, **25**, 383–399.

COX, D. R. (1968). Notes on some aspects of regression analysis. *J. R. Statist. Soc.* A, **131**, 265–279.

—— (1970). *Analysis of Binary Data*. London: Methuen.

DEMPSTER, A. P. (1971). An overview of multivariate data analysis. *J. Multivar. Anal.*, **1**, 316–346.

DYKE, G. V. and PATTERSON, H. D. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics*, **8**, 1–12.

FINNEY, D. J. (1952). *Probit Analysis*. 2nd ed. Cambridge: University Press.

FISHER, R. A. (1949). A biological assay of tuberculosis. *Biometrics*, **5**, 300–316.

GOOD, I. J. (1967). Analysis of log-likelihood ratios, "ANOΛ". (A contribution to the discussion of a paper on least squares by F. J. Anscombe.) *J. R. Statist. Soc.* B, **29**, 39–42.

IRELAND, C. T. and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, **55**, 179–188.

IRWIN, J. O. (1949). A note on the subdivision of χ^2 into components. *Biometrika*, **36**, 130–134.

KENDALL, M. G. and STUART, A. (1967). *The Advanced Theory of Statistics*, 2nd ed, Vol. II. London: Griffin.

KIMBALL, A. W. (1954). Short-cut formulas for the exact partition of χ^2 in contingency tables. *Biometrics*, **10**, 452–458.

KU, H. H., VARNER, R. N. and KULLBACK, S. (1971). On the analysis of multidimensional contingency tables. *J. Amer. Statist. Ass.*, **66**, 55–64.

KULLBACK, S. (1968). *Information Theory and Statistics*. 2nd ed. New York: Dover.

LANCASTER, H. O. (1949). The derivation and partition of χ^2 in certain discrete distributions. *Biometrika*, **36**, 117–128.

—— (1950). The exact partition of χ^2 and its application to the problem of the pooling of small expectations. *Biometrika*, **37**, 267–270.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. London: Wiley.

MAXWELL, A. E. (1961). *Analysing Qualitative Data*. London: Methuen.

NELDER, J. A. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, **22**, 128–141.

—— (1968). Weighted regression, quantal response data and inverse polynomials. *Biometrics*, **24**, 979–985.

PATIL, G. P. and SHORROCK, R. (1965). On certain properties of the exponential-type families. *J. R. Statist. Soc.* B, **27**, 94–99.

SIMPSON, C. H. (1951). The interpretation of interaction in contingency tables. *J. R. Statist. Soc.* B, **13**, 238–241.

YATES, F. (1940). The recovery of inter-block information in balanced incomplete block designs. *Ann. Eugen.*, **10**, 317–325.

—— (1948). The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, **35**, 176–181.

—— (1970). *Experimental Design: Selected Papers*. London: Griffin.

APPENDIX

We can sometimes simplify the expression for deviance. In what follows, summation is to be taken to be over the observations unless otherwise stated.

Theorem. If either (a) $Y=\mu^{\alpha}$ or (b) $Y=\log\mu$ and a constant term is being fitted in the model (say $Y=\sum\beta_i x_i$ where x_0 takes the constant value 1) then $\sum\{(z-\hat{\mu})\,\hat{\mu}/\hat{V}\}=0$, where $\hat{\mu}$ and $\hat{V}$ are the maximum likelihood estimates of μ and V.

Proof. For case (b)

$$\frac{\partial L}{\partial\beta_0}=\sum\frac{(z-\mu)\,\mu}{V}=0.$$

In case (a), we have

$$\frac{\partial L}{\partial\beta_i}=\frac{z-\mu}{V}\frac{d\mu}{dY}x_i=\sum\frac{z-\mu}{V}\frac{\mu}{\theta Y}x_i.$$

Hence

$$\sum\frac{(z-\hat{\mu})\,\hat{\mu}}{\hat{V}\hat{Y}}x_i=0.$$

Then

$$\sum\frac{(z-\hat{\mu})\,\hat{\mu}}{\hat{V}}=\sum\frac{\hat{Y}(z-\hat{\mu})\,\hat{\mu}}{\hat{V}\hat{Y}}$$

$$=\sum\sum_i\hat{\beta}_i x_i\frac{(z-\hat{\mu})\,\hat{\mu}}{\hat{V}\hat{Y}}=\sum_i\left\{\beta_i\sum\frac{(z-\hat{\mu})\,\hat{\mu}}{\hat{V}\hat{Y}}x_i\right\}=0.$$

This completes the proof. When the conditions of this theorem are satisfied, we have, for the Poisson distribution

$$\sum z-\hat{\mu}=0$$

and for the gamma distribution

$$\sum\frac{(z-\hat{\mu})}{\hat{\mu}}=0.$$

Then the deviances for these distributions simplify as follows:

$$\text{Poisson}\quad 2\sum z\ln(z/\mu),$$

$$\text{Gamma}\quad -2\sum\ln(z/\mu).$$