Generalized Linear Models Specified in Terms of Constraints

By R. W. M. Wedderburn

Rothamsted Experimental Station

[Received December 1972. Final revision December 1973]

Summary

A modification of the method of Nelder and Wedderburn (1972) is given for fitting models with the same error distributions as discussed there but with the systematic part of the models specified in terms of constraints. It is possible to fit these by the method described by Nelder and Wedderburn using iterative weighted regression, but it turns out to be simpler to replace each regression calculation of that method by another one with the property that the fitted values of one regression calculation are the residuals of the other and vice versa. The method is applied to testing for marginal homogeneity in contingency tables.

Keywords: CONSTRAINED ESTIMATION; CONTINGENCY TABLES; GENERALIZED LINEAR MODELS; LINEAR MODELS; MARGINAL HOMOGENEITY; MAXIMUM LIKELIHOOD; REGRESSION; WEIGHTED LEAST SQUARES

Introduction

In 1972 Nelder and Wedderburn defined a class of models for which maximum likelihood estimates can be obtained by an iterative procedure in which each iteration involved calculating a weighted linear regression. These models had a random component specifying the distribution of the observations. Several distributions were possible including the normal, Poisson, binomial and gamma distributions. The systematic part of the models specified that some function of the means was linear in a set of parameters. This result generalized the results of Nelder (1968) and the well-known method for obtaining maximum likelihood estimates in probit analysis. Other examples were various models involving contingency tables where the effects were additive on a log scale, the inverse polynomial models of Nelder (1966) and models involving sums of squares which had a χ^2 or gamma distribution.

This paper considers models in which the systematic component of the model is defined by a set of linear constraints. After considering the case of normal linear models specified in this way, the extension to more general models comes naturally. As an example the method is applied to testing for marginal homogeneity in contingency tables.

1. Least-squares Fitting of Models specified in Terms of Constraints

Suppose we have an n-dimensional vector of observations $\mathbf{y}$ from the model

$$E(\mathbf{y}) = \mathbf{Y} \qquad (1)$$

where $\mathbf{LY} = \mathbf{0}$, $\mathbf{L}$ being a $p \times n$ matrix. ($E(\mathbf{y})$ denotes the expectation of $\mathbf{y}$.)

Then suppose that we want to fit this model using weighted least squares, choosing $\hat{\mathbf{Y}}$ to minimize $(\mathbf{y}-\hat{\mathbf{Y}})'\mathbf{W}(\mathbf{y}-\hat{\mathbf{Y}})$ where $\mathbf{W}$ is a given symmetric and positive definite matrix. In the applications described below, $\mathbf{W}$ will be diagonal.

Let $M(\mathbf{A})$ denote the column space of a matrix $\mathbf{A}$. Suppose that $\mathbf{X}$ is a matrix with dimensions $n \times q$ such that $M(\mathbf{X})$ is the orthogonal complement of $M(\mathbf{L}')$ in n-dimensional Euclidean space. Then $\mathbf{LY} = \mathbf{0}$ implies that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ for some q-dimensional vector $\boldsymbol{\beta}$. (Here we have $p+q \geqslant n$ with equality if both the rows of $\mathbf{L}$ and the columns of $\mathbf{X}$ are linearly independent.) So (1) may be written

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}. \tag{2}$$

We know how to fit (2) by weighted least squares obtaining $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{y}$. (Here $\mathbf{A}^{-}$ denotes a generalized inverse of $\mathbf{A}$, i.e. any matrix satisfying $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$.) $\hat{\mathbf{Y}}$ is, of course, unique even though $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}$ may not be.

However, if p is small and n is large, then q is also large, and we have a large matrix to invert; also determining $\mathbf{X}$ explicitly may be quite difficult. Fortunately, we can obtain $\hat{\mathbf{Y}}$ much more easily.

First, note that $\mathbf{W}$ defines an inner product on R^n. For the rest of this section the term "W-orthogonal" will mean orthogonal with respect to this inner product (i.e. $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}'\mathbf{W}\mathbf{y}$). Now, since $\mathbf{W}$ is non-singular, it is easily seen that $M(\mathbf{W}^{-1}\mathbf{L}')$ and $M(\mathbf{X})$ are W-orthogonal complements.

Let $\mathbf{r}$ be the vector of residuals obtained by fitting (2) using $\mathbf{W}$ as weight matrix. Then

$$\mathbf{y} = \hat{\mathbf{Y}} + \mathbf{r}.$$

Now $\hat{\mathbf{Y}}$ is the W-orthogonal projection of $\mathbf{y}$ into $M(\mathbf{X})$. Thus $\mathbf{r}$ is the projection of $\mathbf{y}$ into the W-orthogonal complement of $M(\mathbf{X})$ which is $M(\mathbf{W}^{-1}\mathbf{L})$. It follows that $\mathbf{r}$ is the vector of fitted values that would be obtained from a least-squares fit of

$$E(\mathbf{y}) = \mathbf{W}^{-1}\mathbf{L}\boldsymbol{\gamma} \tag{3}$$

using $\mathbf{W}$ as the weight matrix. In other words, the fitted values for (2) will be the residuals for (3) and vice versa. Fitting the regression (3) is likely to be quite easy if p is small and $\mathbf{W}$ is diagonal.

In fact, the following result has now been proved.

Theorem. If

(i) $\mathbf{X}$ is a matrix of maximal rank such that $\mathbf{L}'\mathbf{X} = 0$,
(ii) $\mathbf{W}$ is a symmetric positive definite matrix and
(iii) $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are least-squares estimates for $(\boldsymbol{\beta})$ and $(\boldsymbol{\gamma})$ in the equations $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $E(\mathbf{y}) = \mathbf{W}^{-1}\mathbf{L}\boldsymbol{\gamma}$ with $\mathbf{W}$ used as the weight matrix, then $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{W}^{-1}\mathbf{L}'\hat{\boldsymbol{\gamma}}$.

It cannot be claimed that this result is entirely new although it is difficult to find a satisfactory statement. Essentially the same result, though expressed and derived in a very different way, can be found in Chapter VII of Brunt (1917).

2. Extension to Generalized Linear Models

The class of models discussed by Nelder and Wedderburn (1972) took the form

$$E(\mathbf{z}) = \boldsymbol{\mu}, \tag{4}$$

where $\mu_i = f(Y_i)$ and $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$.

The distribution of $\mathbf{z}$ may come from a one-parameter exponential family of distributions or a family with densities of the form

$$\pi(z;\ \theta, \phi) = \exp\left[\alpha(\phi)\{z\theta - g(\theta) + h(z)\} + \beta(\phi, z)\right], \tag{5}$$

where ϕ is a fixed nuisance parameter, and θ varies from one observation to another according to the model (4), the relation between θ and μ being $\mu = g'(\theta)$ (Section (1.1) of Nelder and Wedderburn, 1972). The possible distributions for z_i include the normal distribution, possibly with unknown variance, binomial distribution, the Poisson distribution and the gamma distribution, possibly with unknown coefficient of variation.

It was shown that we can write $\text{var}(z)$ in the form $V(\mu)/\alpha(\phi)$ or simply $V(\mu)$ if there is no nuisance parameter, where $V(\mu) = g''(\theta)$. Then each iteration in the fitting of the model by maximum likelihood consists of regressing the vector whose components are

$$y = Y + (z-\mu)\frac{dY}{d\mu} \tag{6}$$

on **X**, and using weights

$$w = \left(\frac{d\mu}{dY}\right)^2 \bigg/ V(\mu). \tag{7}$$

Here μ and Y stand for the current approximation to $\hat{\mu}$ and $\hat{Y}$ and the fitted values of the regression provide a new value for Y from which a new value for μ can be calculated. We generally start by taking $\mu = \mathbf{z}$.

Suppose now that we have the same distributional assumptions for **z**, but the model is expressed in the form

$$E(\mathbf{z}) = \boldsymbol{\mu}, \quad \text{where } \mu_i = f(Y_i) \text{ and } \mathbf{LY} = \mathbf{0}. \tag{8}$$

Following the argument of Section 1 we can find **X** so that the model takes the form (4). Then the iterative method already described can be used, and each step of the iteration is equivalent to fitting the regression model (2). Clearly we may fit (3) instead, and then use the residuals as the new value for Y. The algorithm then takes the following form:

(a) Set $\boldsymbol{\mu} = \mathbf{z}$ and calculate **Y** from $\boldsymbol{\mu}$. Set $\mathbf{y} = \mathbf{Y}$.

(b) Calculate the diagonal matrix **W** using (7).

(c) Regress **y** on $\mathbf{W}^{-1}\mathbf{L}'$ using **W** as weight matrix. Set **Y** = residuals and calculate $\boldsymbol{\mu}$ from **Y**. If the process has gone far enough stop.

(d) Calculate **y** from (6) and go to step (b).

If we have inhomogeneous constraints for the form

$$\mathbf{LY} = \mathbf{c}$$

then we simply have to choose any vector **a** such that

$$\mathbf{La} = \mathbf{c}$$

and redefine f in (8) so that **Y** is replaced by $\mathbf{Y}-\mathbf{a}$. The constraints then become homogeneous. This means that the function f will be different for each observation, but it should be noted that although the algorithm has been described as if f of (8) and g and h of (5) were the same for each observation, there is no need for them to be so.

3. APPLICATION—MARGINAL HOMOGENEITY IN CONTINGENCY TABLES

Several authors have considered testing whether the two margins of a square contingency table may be considered to have equal expectation, e.g. Stuart (1955) and Ireland, Ku and Kullback (1969). Stuart (1955) says that the likelihood-ratio principle yields an intractable result. Kullback (1971a) considered the corresponding problem for multidimensional tables.

It turns out that the method described above can be used to provide maximum likelihood estimates of the cell frequencies p_{ij} subject to the constraints

$$\sum_i p_{ij} = \sum_i p_{ji} \quad \text{for } j = 1, \ldots, n.$$

Let the observations be n_{ij}. As pointed out in Nelder and Wedderburn (1972), for the purposes of maximum likelihood estimation we can treat a sample from a multinomial distribution as if it consisted of independent Poisson observations. This is most easily seen as follows: suppose we have a sample $n_1, \ldots, n_k$ from a multinomial distribution with parameters $p_1, \ldots, p_k$; the log-likelihood is

$$L(p_1, \ldots, p_k) = \sum_i n_i \ln p_i.$$

If we regard the observations as coming from Poisson distributions with mean mp_i, we obtain a log-likelihood

$$L^*(p_1, \ldots, p_k, m) = \sum n_i \ln m - m + L(p_1, \ldots, p_k).$$

The term $\sum n_i \ln m - m$ does not involve $p_1, \ldots, p_k$ and so likelihood ratio tests of hypotheses concerning the p's using L^* will give the same results as tests using L.

If the expected cell frequencies are μ_{ij}, the hypothesis of marginal homogeneity may be expressed as

$$\sum_i \mu_{ij} = \sum_i \mu_{ji} \quad \text{for } j = 1, \ldots, n.$$

Only $n-1$ of these constraints are linearly independent. Using the notation already developed for generalized linear models, we call the observed frequencies z_{ij}. We have $Y_{ij} = \mu_{ij}$, so that the y-variate for each iteration is just the set of z_{ij}'s, and $w_{ij} = 1/\mu_{ij}$. The kth constraint may be written $\sum_{ij} l_{ij}^{(k)} \mu_{ij} = 0$ where $l_{ij}^{(k)} = \delta_{ik} - \delta_{jk}$ where δ_{ik}, etc. are Kronecker δ symbols.

Then we use as independent variates

$$x_{ij}^{(k)} = \mu_{ij}(\delta_{ik} - \delta_{jk}) \quad \text{for } k = 1, 2, \ldots, n-1$$

and the residuals provide the next approximation to μ_{ij}.

Example 1

Stuart (1955) applied a test of marginal homogeneity to the data of Table 1.

Stuart obtained a value for χ^2 of 11·96 with 3 degrees of freedom, Ireland, Ku and Kullback (1969) using several methods obtained 11·998, 12·010 and 11·978. The χ^2 calculated from the likelihood ratio was called the "deviance" by Nelder and Wedderburn; the value obtained using the method of this paper was 11·986. Of course, with quite large numbers in the table we would expect all the methods, which are asymptotically equivalent, to give similar answers. The values of $\hat{\mu}_{ij}$ are given along with the data in Table 1. We notice that the observed values are consistently larger than the fitted values above the diagonal of the table and consistently less below the diagonal. It seems that left eye vision tended to be weaker than right eye vision. We inevitably find that $\hat{\mu}_{ii} = z_{ii}$. This is not the case in Table 6.3 of Ireland, Ku and Kullback and Stuart's method does not provide expected frequencies at all.

TABLE 1

Unaided distance vision of 7,477 *women aged* 30–39 *with fitted values* ($\hat{\mu}_{ij}$) *in brackets*

Left eye / *Right eye*	*Highest grade*	*Second grade*	*Third grade*	*Lowest grade*	*Total*
Highest grade	1,520 (1,520·0)	266 (252·5)	124 (111·8)	66 (57·0)	1,976 (1,941·3)
Second grade	234 (247·2)	1,512 (1,512·0)	432 (409·4)	78 (70·6)	2,256 (2,239·2)
Third grade	117 (131·3)	362 (383·1)	1,772 (1,772·0)	205 (195·3)	2,456 (2,481·7)
Fourth grade	36 (42·8)	82 (91·6)	179 (188·4)	492 (492·0)	789 (814·8)
Total	1,907 (1,941·3)	2,222 (2,239·2)	2,507 (2,481·7)	841 (814·8)	7,477 (7,477·0)

Example 2. *A* 2^4 *contingency table*

Kullback (1971b) considered the data in Table 2 which shows the distribution of the sexes in the first four births in 36,536 families. He gave a test to determine whether the sex ratios in the first four births are equal. The constraints here are

$$\sum_{jkl}\mu_{ijkl} = \sum_{jkl}\mu_{jikl} = \sum_{jkl}\mu_{jkil} = \sum_{jkl}\mu_{jkli}.$$

As in the previous example we have $\boldsymbol{\mu} = \mathbf{Y}$.

TABLE 2

Distribution of sexes among first four birth orders

		k				
		M		*F*		
		l				
i	*j*	*M*	*F*	*M*	*F*	
M	M	2,574	2,469	2,401	2,313	9,757
	F	2,478	2,289	2,329	2,121	9,217
F	M	2,340	2,258	2,276	2,209	9,083
	F	2,253	2,084	2,107	2,035	8,479
		9,645	9,100	9,113	8,678	36,536

A likelihood ratio test gave a deviance of 3·656 with 3 d.f. This compares with 3·672 produced by Kullback. It seems that the data show little evidence of inequality of the sex ratios in different birth orders.

These calculations were done using the GENSTAT system developed at Rothamsted, which includes all the necessary facilities for iterative regression by allowing the storing of residuals or fitted values, general arithmetic vector operations and looping.

4. CONCLUSION

The method of this paper is an extension of that originally described for generalized linear models. An application has been described, namely that of testing the hypothesis of marginal homogeneity of contingency tables. In the absence of a convenient method for applying maximum likelihood to this problem, other methods have proliferated. No doubt other applications of the method (such as one concerning variance component estimation to be described by G. N. Wilkinson elsewhere) will be found.

REFERENCES

BRUNT, D. (1917). *The Combination of Observations*. Cambridge: University Press.

IRELAND, C. T., KU, H. H. and KULLBACK, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *J. Amer. Statist. Ass.*, **64**, 1323–1341.

KULLBACK, S. (1971a). Marginal homogeneity of multidimensional contingency tables. *Ann. Math. Statist.*, **42**, 594–606.

—— (1971b). The homogeneity of the sex ratio of adjacent sibs in human families. *Biometrics*, **27**, 452–457.

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

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

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.

STUART, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412–416.