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

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

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

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

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

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

Now

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

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

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

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

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

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 $W d\eta/d\mu$ 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).