AMS 274: Generalized Linear Models
Fall 2016

1. Definitions

Exponential dispersion family of distributions

Central to the development of generalized linear models (GLMs) is the exponential dispersion family of distributions as it defines the random component for a GLM, and extends (normal) linear models, since it includes the normal distribution as a special case.

Consider a (univariate) random variable Y that takes values in a set S that can be countable or uncountable, that is, Y can be discrete or continuous. We say that the distribution of Y belongs to the exponential dispersion family of distributions if its probability density/mass function has the form

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \qquad (1)$$

where the parameter θ takes values in a subset of $\mathbb{R}$, the *dispersion* parameter ϕ takes values in a subset of $\mathbb{R}^+$, the support S of the distribution does not depend on the parameters θ and ϕ, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot,\cdot)$ are specified functions of ϕ, θ, and (y, ϕ), respectively. For specific choices of $a(\phi)$ (> 0) and $b(\theta)$, $c(y, \phi)$ is a normalizing function that ensures that (1) defines a valid density or mass function. The exponential dispersion family includes the Binomial (for known number of trials) and Poisson distributions ($a(\phi) = 1$ for both) as well as the normal, gamma, and inverse Gaussian distributions.

Note that for ϕ known, (1) is a special case of the one-parameter exponential family of distributions (with natural parameter θ). It may or may not be a two-parameter exponential family when both θ and ϕ are unknown (it depends on the actual form of $c(y, \phi)$). The term exponential dispersion family reflects this partly exponential form of (1) as well as the important role played by the dispersion parameter ϕ. The exponential dispersion family allows more flexibility than the one-parameter exponential family through dispersion parameter ϕ. At the same time, its form is more convenient for estimation than the two-parameter exponential family.

Using standard results for expectations of derivatives of log-likelihood functions (see the Appendix), we obtain

$$\mathrm{E}(Y \mid \theta, \phi) = b'(\theta)$$

and

$$\mathrm{Var}(Y \mid \theta, \phi) = a(\phi) b''(\theta). \qquad (2)$$

(Here, $b'(\theta) = db(\theta)/d\theta$ and $b''(\theta) = d^2b(\theta)/d\theta^2$.) Note the specific mean-variance relationship, $\mathrm{Var}(Y) = a(\phi) d\mathrm{E}(Y)/d\theta$, implied by the exponential dispersion family. This relationship, along with the inability of the family to model moments of order higher than 2, indicate some of the implicit restrictions for standard GLMs. The function $a(\phi)$ is typically of the form $a(\phi) = \phi$ or $a(\phi) = \phi/w$, where w is a known *weight* that depends on the observation, clarifying further, through expression (2), the role of ϕ as a dispersion parameter.

The function $b''(\theta)$ is typically referred to as the *variance function* of the family. It can be shown that it characterizes the member of the exponential dispersion family. The fact that the normal distribution has a variance function that does not depend on its mean (specifically, $b''(\theta) = 1$ for the normal distribution) emphasizes its distinguishing role within the class of exponential dispersion models, allowing the formulation of the standard normal linear model with constant variance.

For a systematic treatment of exponential dispersion models, including multivariate generalizations of (1), see Jørgensen (1987).

GLM structure

For the definition of the standard class of GLMs (Nelder and Wedderburn, 1972), we consider the following three components for the model structure.

1. Random component: The response observations y_i, $i = 1, ..., n$, are assumed to be realizations of random variables Y_i that are independent and follow a distribution that is a member of the exponential dispersion family with common dispersion parameter ϕ. That is,

$$Y_i \overset{ind.}{\sim} f(y_i \mid \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right), \quad i = 1, ..., n, \tag{3}$$

with means $\mu_i = \mathrm{E}(Y_i \mid \theta_i, \phi) = b'(\theta_i)$ and variances $\mathrm{Var}(Y_i \mid \theta_i, \phi) = a_i(\phi)b''(\theta_i)$.

2. Systematic component: This is the linear predictor, familiar from normal linear regression modeling. Let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ denote the values, associated with y_i, from a $p \times 1$ vector of (known) explanatory variables (either continuous or categorical). The $n \times p$ design matrix $\boldsymbol{X} = (\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T)^T$ collects the values from all explanatory variables for all observations. Denoting by $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ the vector of regression coefficients, the linear predictor is

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, ..., n.$$

3. Link function: The link function g is a transformation of the mean that addresses problems of scaling, since the mean of the exponential dispersion family does not necessarily take values in $\mathbb{R}$. The link function specifies the relationship between the mean of the ith response and the associated linear predictor and hence connects (links) the random and systematic components of the model,

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad i = 1, ..., n. \tag{4}$$

In the definition of standard GLMs, g is assumed to be a monotonic and differentiable function (an invertible g facilitates interpretation and maximum likelihood fitting for GLMs). It is typically a specified function, although link functions that depend on a number of parameters (which can be estimated from the data) have been studied in the literature. Nonparametric modeling, including Bayesian nonparametric approaches, for the link function has also been studied.

Note from (4) that $\mu_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$, that is, a linear regression for the means on a transformed scale given through the link function. An expression for the θ_i in terms of the $\boldsymbol{x}_i^T\boldsymbol{\beta}$ also emerges noting that $\mu_i = b'(\theta_i)$. In fact, the *canonical link* is given by $g(\mu_i) = \theta_i(\mu_i)$, under which $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

2. Least squares methods for linear and non-linear regression

Linear regression

Consider the familiar form of the linear regression model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i, \quad i = 1, ..., n, \tag{5}$$

or $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ in matrix-vector notation, where $\boldsymbol{y} = (y_1, ..., y_n)^T$ and $\boldsymbol{e} = (e_1, ..., e_n)^T$.

Point estimation for the regression coefficients is possible with assumptions only for the first two moments of the response distribution. Under the standard assumptions, $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{e}) = \sigma^2 I_n$, the *least squares* (LS) estimates $\hat{\boldsymbol{\beta}}_{LS}$ arise by minimizing the function $\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, yielding $\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ (provided $\boldsymbol{X}$ is of full rank).

Under the more general assumption $\mathrm{Var}(\boldsymbol{e}) = \sigma^2 \Sigma$, where Σ is a known $n \times n$ matrix, minimization of the function $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ yields the *generalized least squares* estimates $\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T \Sigma^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \Sigma^{-1} \boldsymbol{y}$ (provided all required matrix inverses exist).

Introduction of the matrix Σ allows for correlated errors and different error variances. The special case $\Sigma = \mathrm{diag}(w_1^{-1}, ..., w_n^{-1})$, where w_i are known *weights*, leads to the *weighted least squares* (WLS) estimates through minimization of the function $\sum_{i=1}^n w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$. Here, although the familiar assumption of uncorrelated errors is retained, observations with large weight (i.e., small variance) contribute more to the function that is minimized. Note that, defining $C = \mathrm{diag}(\sqrt{w_1}, ..., \sqrt{w_n})$, the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, with $\mathrm{Var}(\boldsymbol{e}) = \sigma^2 \mathrm{diag}(w_1^{-1}, ..., w_n^{-1})$, can be transformed to the linear model $C\boldsymbol{y} = C\boldsymbol{X}\boldsymbol{\beta} + C\boldsymbol{e}$, where $\mathrm{Var}(C\boldsymbol{e}) = \sigma^2 I_n$. Hence, the WLS estimates can be obtained using the LS method for the transformed model.

Note that in the estimation approaches discussed above the assumptions for the response distribution involve only its first two moments. Hence any inference other than point estimation is essentially not possible. For further inference (e.g., interval estimation or hypothesis testing) one has to resort to asymptotic results or perhaps bootstrap methods.

Model (5) can be fully parameterized by specifying a (parametric) response distribution. It is well known that with a normal response distribution (i.e., $\boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 I_n)$) the maximum likelihood estimates for $\boldsymbol{\beta}$ are the same with the LS estimates $\hat{\boldsymbol{\beta}}_{LS}$. In a similar fashion, other optimization methods can be justified as maximum likelihood approaches under specific response distributions.

Non-linear regression

In certain applications, we may anticipate a non-linear relationship between the response and the explanatory variables. Denoting by $h(\boldsymbol{x}, \boldsymbol{\beta})$ a generic (non-linear) regression function defined in terms of the vector $\boldsymbol{\beta}$ of regression coefficients, the natural extension of model (5) is

$$y_i = h(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i, \quad i = 1, ..., n. \tag{6}$$

The errors are typically assumed uncorrelated, possibly with different variances, e.g., $\mathrm{Var}(e_i) = \sigma^2 / w_i$, for known weights w_i. Again, a likelihood approach to estimation can be taken if a specific parametric response distribution is employed.

To proceed with assumptions only for the first two moments, an extension of the WLS technique provides a possible approach to estimation for $\boldsymbol{\beta}$. The estimates of $\boldsymbol{\beta}$ will now minimize the function

$$I(\boldsymbol{\beta}) = \sum_{i=1}^n (\mathrm{Var}(Y_i))^{-1} (y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta}))^2 \tag{7}$$

which is similar to the one used in the LS or WLS approach when the $\mathrm{Var}(Y_i)$ are constant or of the form σ^2/w_i, respectively. However, here the non-linearity of the regression function introduces complications. Note that, in general, $\mathrm{Var}(Y_i)$ can depend on $\boldsymbol{\beta}$ adding to the complexity of the estimation procedure. In what follows, we assume that $\mathrm{Var}(Y_i)$ is not a function of $\boldsymbol{\beta}$.

The estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, arising from the minimization of (7), must satisfy the system of equations (typically called the normal equations)

$$\frac{\partial I(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} (\mathrm{Var}(Y_i))^{-1}(y_i - h(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})) \left[\frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_j}\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0, \quad j = 1, ..., p.$$

The term $\partial h(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial \beta_j$ shows that the normal equations are, in general, non-linear due to the non-linearity of h (note that $\partial h(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial \beta_j = x_{ij}$, free of $\boldsymbol{\beta}$, for the linear regression model).

To address the lack of closed form solutions, numerical (iterative) methods are needed. **Iterative weighted least squares** (IWLS) is one such method that utilizes an approximate linearization of h and applies WLS in an iterative fashion. Specifically to implement the IWLS technique:

(i) Start with initial values $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_p^0)$.

(ii) Approximate h with a linear form (linear in the β_j). For example, a first-order Taylor series expansion of $h(\boldsymbol{x}_i, \boldsymbol{\beta})$ about $\boldsymbol{\beta}^0$ yields

$$h(\boldsymbol{x}_i, \boldsymbol{\beta}) \approx h^*(\boldsymbol{x}_i, \boldsymbol{\beta}) = h(\boldsymbol{x}_i, \boldsymbol{\beta}^0) + \sum_{j=1}^{p} \left[\frac{\partial h(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_j}\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}^0} (\beta_j - \beta_j^0). \tag{8}$$

If we use $h^*(\boldsymbol{x}_i, \boldsymbol{\beta})$ from (8) to approximate $h(\boldsymbol{x}_i, \boldsymbol{\beta})$, the non-linear regression model (6) is approximated by the linear regression model

$$u_i^0 = \sum_{j=1}^{p} z_{ij}^0 \delta_j^0 + e_i, \quad i = 1, ..., n, \tag{9}$$

with transformed responses $u_i^0 = y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta}^0)$, explanatory variables $z_{ij}^0 = (\partial h(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial \beta_j)_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}$, and regression coefficients $\delta_j^0 = \beta_j - \beta_j^0$. Note that the values of both the response and explanatory variables in (9) depend on the initial values $\boldsymbol{\beta}^0$.

(iii) Solve the linear system that results from (9) to obtain an estimate $\hat{\boldsymbol{\delta}}^0 = (\hat{\delta}_1^0, ..., \hat{\delta}_p^0)$ of $\boldsymbol{\delta}^0 = (\delta_1^0, ..., \delta_p^0)$. The estimate is obtained by minimizing (using either LS or WLS)

$$\sum_{i=1}^{n} (\mathrm{Var}(Y_i))^{-1} \left(y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta}^0) - \sum_{j=1}^{p} z_{ij}^0 \delta_j^0 \right)^2 \tag{10}$$

with respect to the δ_j^0, $j = 1, ..., p$. For example, under the assumption $\mathrm{Var}(e_i) = \sigma^2$ in (6) and denoting by $\boldsymbol{Z}^0$ the $n \times p$ matrix defined by the z_{ij}^0 and $\boldsymbol{u}^0 = (u_1^0, ..., u_n^0)^T$, the estimate $\hat{\boldsymbol{\delta}}^0 = (\boldsymbol{Z}^{0T}\boldsymbol{Z}^0)^{-1}\boldsymbol{Z}^{0T}\boldsymbol{u}^0$ (provided $\boldsymbol{Z}^0$ is of full rank). Note the difference between expressions (7) and

(10). The former involves the non-linear regression function h and its solution would yield the actual WLS estimates of $\boldsymbol{\beta}$. However, we work with the latter that provides approximate (revised) estimates for $\boldsymbol{\beta}$.

(iv) Since $\boldsymbol{\delta}^0 = \boldsymbol{\beta} - \boldsymbol{\beta}^0$, we can write $\hat{\boldsymbol{\delta}}^0 = \boldsymbol{\beta}^1 - \boldsymbol{\beta}^0$ setting $\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 + \hat{\boldsymbol{\delta}}$ as the current (revised) estimate of $\boldsymbol{\beta}$.

Steps (ii) to (iv) are repeated until the resulting sequence $\{\boldsymbol{\beta}^k\}$ *converges.* Assuming, again, $\text{Var}(e_i) = \sigma^2$ and in obvious extension of the previous notation, the iteration of the IWLS algorithm can be expressed as

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k + (\boldsymbol{Z}^{kT}\boldsymbol{Z}^k)^{-1}\boldsymbol{Z}^{kT}\boldsymbol{u}^k.$$

Note that both $\boldsymbol{u}^k$ and $\boldsymbol{Z}^k$ depend on the current estimate $\boldsymbol{\beta}^k$ clarifying the iterative nature of the IWLS approach.

Convergence can be slow, or might not even be obtained, depending on the form of h and the initial values. Under conditions on h and for certain response distributions, IWLS estimates approximate the associated maximum likelihood estimates; see, e.g., Charnes, Frome and Yu (1976) for a result involving the one-parameter exponential family of distributions. Refer to Draper and Smith (1981) for more details as well as discussion of other numerical estimation methods for non-linear regression models.

3. Newton-Raphson and scoring methods for numerical maximization

Let $t(\theta)$ be a real-valued function that we wish to maximize with respect to θ. If there is no closed form solution to the problem, numerical (iterative) approaches emerge as useful options. Of course, this is a numerical analysis problem, the important statistical application being to maximum likelihood estimation where $t(\theta) = l(\theta; \boldsymbol{y}) = \log L(\theta; \boldsymbol{y}) = \log \prod_{i=1}^{n} f(y_i \mid \theta)$, the log-likelihood function for θ based on the observed data $\boldsymbol{y} = (y_1, ..., y_n)$. Assume first that θ is one dimensional.

The **Newton-Raphson algorithm** (or **Newton's method**) is an iterative approach that employs a quadratic Taylor series approximation of $t(\theta)$. Specifically, for a starting value θ^0, $t(\theta)$ is approximated by a second-order Taylor series expansion about θ^0,

$$t(\theta) \approx t^*(\theta) = t(\theta^0) + t'(\theta^0)(\theta - \theta^0) + 0.5t''(\theta^0)(\theta - \theta^0)^2,$$

where $t'(\theta^0) = (dt(\theta)/d\theta)_{\theta=\theta^0}$ and $t''(\theta^0) = (d^2t(\theta)/d\theta^2)_{\theta=\theta^0}$. Next, the new iterate θ^1 is determined from $dt^*(\theta)/d\theta = 0$, hence

$$\theta^1 = \theta^0 - \frac{t'(\theta^0)}{t''(\theta^0)}, \tag{11}$$

and the procedure is repeated until *convergence.*

In the context of maximum likelihood estimation, $t'(\theta) = \partial l(\theta; \boldsymbol{y})/\partial\theta = U(\theta; \boldsymbol{y})$, the *score function,* and $t''(\theta) = \partial^2 l(\theta;\boldsymbol{y})/\partial\theta^2 = \partial U(\theta;\boldsymbol{y})/\partial\theta \equiv U'(\theta;\boldsymbol{y})$ and the iteration of the Newton-Raphson algorithm becomes

$$\theta^{k+1} = \theta^k - \frac{U(\theta^k;\boldsymbol{y})}{U'(\theta^k;\boldsymbol{y})}.$$

The **scoring method** (or **Fisher's scoring method**) is a variant of the Newton-Raphson method that replaces $U'(\theta;\boldsymbol{y})$ with $\mathrm{E}(U'(\theta;\boldsymbol{Y}))$, where the expectation is taken with respect to the distribution of $\boldsymbol{Y} = (Y_1, ..., Y_n)$. Recall that, under regularity conditions, we have $\mathrm{E}(-U') = \mathrm{Var}(U) = J(\theta)$, the (expected) Fisher information. Hence the iteration of the scoring method is given by

$$\theta^{k+1} = \theta^k + \frac{U(\theta^k;\boldsymbol{y})}{J(\theta^k)}.$$

Note that, under regularity conditions on the underlying distribution with density/mass function $f(\cdot \mid \theta)$, the maximum likelihood estimate $\hat{\theta}$ of θ is asymptotically normal with variance inversely related to $J(\theta)$, $\mathrm{Var}(\hat{\theta}) \approx 1/J(\theta)$. The curvature of $l(\theta;\boldsymbol{y})$ for $\theta \in (\hat{\theta}-\epsilon, \hat{\theta}+\epsilon)$ specifies the precision of $\hat{\theta}$ and this curvature depends on the rate of change of $\partial l(\theta;\boldsymbol{y})/\partial\theta = U(\theta;\boldsymbol{y})$, i.e., on $U'(\theta;\boldsymbol{y})$ and hence also on $\mathrm{E}(U'(\theta;\boldsymbol{y}))$. A flat log-likelihood $l(\theta;\boldsymbol{y})$ implies an imprecise estimate $\hat{\theta}$ and this, in turn, relates to a large standard error $\sqrt{\mathrm{Var}(\hat{\theta})}$.

Turning to a multiparameter setting, $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$, both approaches can be extended, using second-order Taylor series expansion for the real-valued function $t(\boldsymbol{\theta})$, which now has a multidimensional argument, and replacing the first-order derivative $t'(\theta)$ with the vector $t'(\boldsymbol{\theta}) = (\partial t(\boldsymbol{\theta})/\partial\theta_1, ..., \partial t(\boldsymbol{\theta})/\partial\theta_m)^T$ and the second-order derivative $t''(\theta)$ with the $m \times m$ matrix $t''(\boldsymbol{\theta})$ with elements $\partial^2 t(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j$, for $i, j = 1, ..., m$. The analogous expression to (11) is

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 - \left(t''(\boldsymbol{\theta}^0)\right)^{-1} t'(\boldsymbol{\theta}^0).$$

For maximum likelihood estimation, $t(\boldsymbol{\theta}) = l(\boldsymbol{\theta};\boldsymbol{y})$, $(\partial l(\boldsymbol{\theta};\boldsymbol{y})/\partial\theta_1, ..., \partial l(\boldsymbol{\theta};\boldsymbol{y})/\partial\theta_m)^T = U(\boldsymbol{\theta};\boldsymbol{y})$ is the *score function vector* and the $m \times m$ matrix $I(\boldsymbol{\theta};\boldsymbol{y})$ with elements $-\partial^2 l(\boldsymbol{\theta};\boldsymbol{y})/\partial\theta_i\partial\theta_j$, for $i, j = 1, ..., m$, is the **observed (Fisher) information matrix**. The Newton-Rapshon method uses the observed Fisher information matrix whereas the scoring method is based on the **expected (Fisher) information matrix** $J(\boldsymbol{\theta})$ with elements $\mathrm{E}(-\partial^2 l(\boldsymbol{\theta};\boldsymbol{Y})/\partial\theta_i\partial\theta_j)$, for $i, j = 1, ..., m$. Hence the iteration for the Newton-Raphson algorithm is given by

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \Big(I(\boldsymbol{\theta}^k;\boldsymbol{y})\Big)^{-1} U(\boldsymbol{\theta}^k;\boldsymbol{y})$$

and for the scoring method

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \Big(J(\boldsymbol{\theta}^k)\Big)^{-1} U(\boldsymbol{\theta}^k;\boldsymbol{y}).$$

From a practical point of view, the choice of the method depends on whether obtaining the expected information matrix (or number) is feasible. In complex multiparameter models, the required expectations are typically not available in closed form and, therefore, the Newton-Raphson method is simpler to implement. Note that for the exponential dispersion family (1) (common θ and ϕ for all y_i) we have $U' = \mathrm{E}(U')$. Moreover, to estimate the regression coefficients in a GLM, the scoring method actually results, in general, to simplifications compared to the Newton-Raphson method.

Finally, we note that use of the expected versus observed information matrix has been studied in the literature with emphasis on the resulting theoretical properties of estimates; see, e.g., Efron and Hinkley (1978).

4. Maximum likelihood estimation for GLMs

Consider the GLM setting described in Section 1. The objective is to obtain maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, which, based on the invariance of maximum likelihood estimation, will also yield maximum likelihood estimates $\hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ for the linear predictors, and $\hat{\mu_i} = g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$ for the means. The approach will be based on the scoring method, since, in this context, use of the expected information matrix provides simplified expressions compared to the observed information matrix. We can also express the iterative approach as an IWLS algorithm.

For the scoring method, we need the score function vector $U(\boldsymbol{\beta}; \boldsymbol{y}) = (U_1(\boldsymbol{\beta}; \boldsymbol{y}), ..., U_p(\boldsymbol{\beta}; \boldsymbol{y}))^T$, where $U_j(\boldsymbol{\beta}; \boldsymbol{y}) = \partial l(\boldsymbol{\beta}; \boldsymbol{y})/\partial \beta_j$, for $j = 1, ..., p$, and the expected information matrix $J(\boldsymbol{\beta})$ with elements $\mathrm{E}(-\partial^2 l(\boldsymbol{\beta}; \boldsymbol{Y})/\partial \beta_k \partial \beta_j)$, for $k, j = 1, ..., p$.

Based on (3), the log-likelihood function for $\boldsymbol{\beta}$ is given by

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i \mid \theta_i, \phi) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right) = \sum_{i=1}^{n} l_i,$$

where $l_i \equiv l_i(\boldsymbol{\beta}; y_i)$ is the contribution to the log-likelihood from the ith observation, depending on $\boldsymbol{\beta}$ through the θ_i (recall that $b'(\theta_i) = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$).

Hence the score function vector has elements

$$U_j(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{\partial l_i}{\beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \tag{12}$$

using the chain rule for differentiation. Now, for each of the derivatives in (12) we have

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}$$

$$\frac{d\theta_i}{d\mu_i} = \left(\frac{d\mu_i}{d\theta_i} \right)^{-1} = \left(\frac{db'(\theta_i)}{d\theta_i} \right)^{-1} = \frac{1}{b''(\theta_i)} = \frac{a_i(\phi)}{\mathrm{Var}(Y_i)}$$

$$\frac{d\mu_i}{d\eta_i} = \frac{dg^{-1}(\eta_i)}{d\eta_i},$$

with a form that depends on the link function g, and

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial}{\beta_j} \sum_{\ell=1}^{p} x_{i\ell} \beta_\ell = x_{ij}.$$

Substituting the above expressions in (12), we obtain

$$U_j(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \frac{d\mu_i}{d\eta_i} x_{ij}. \tag{13}$$

Regarding the expected Fisher information matrix, we have

$$\frac{\partial^2 l(\boldsymbol{\beta};\boldsymbol{y})}{\partial\beta_k \partial\beta_j} = \frac{\partial U_j(\boldsymbol{\beta};\boldsymbol{y})}{\partial\beta_k} = \frac{\partial}{\partial\beta_k} \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \frac{d\mu_i}{d\eta_i} x_{ij},$$

using (13), which results in

$$\frac{\partial^2 l(\boldsymbol{\beta};\boldsymbol{y})}{\partial\beta_k \partial\beta_j} = \sum_{i=1}^{n} x_{ij} \left[\frac{d\mu_i}{d\eta_i} \frac{\partial}{\partial\beta_k} \left(\frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \right) + \frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \frac{\partial}{\partial\beta_k} \left(\frac{d\mu_i}{d\eta_i} \right) \right]. \tag{14}$$

Therefore, the (k,j)th element of the expected information matrix $J(\boldsymbol{\beta})$ is given by

$$(J(\boldsymbol{\beta}))_{k,j} = \mathrm{E}\left(-\frac{\partial^2 l(\boldsymbol{\beta};\boldsymbol{y})}{\partial\beta_k \partial\beta_j} \right) = \mathrm{E}\left[-\sum_{i=1}^{n} x_{ij} \frac{d\mu_i}{d\eta_i} \frac{\partial}{\partial\beta_k} \left(\frac{Y_i - \mu_i}{\mathrm{Var}(Y_i)} \right) \right], \tag{15}$$

since the second term in (14) cancels on taking expectations. Expressions (14) and (15) illustrate the fact that the method of scoring results in simplifications compared to the Newton-Raphson method. Moreover,

$$\begin{aligned} \tfrac{\partial}{\partial\beta_k} \left(\tfrac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \right) &= \tfrac{\partial}{\partial\mu_i} \left(\tfrac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \right) \tfrac{d\mu_i}{d\eta_i} \tfrac{\partial\eta_i}{\partial\beta_k} \\ &= \left(-\tfrac{1}{\mathrm{Var}(Y_i)} \right) \tfrac{d\mu_i}{d\eta_i} x_{ik}, \end{aligned}$$

that can be substituted in (15) to yield

$$(J(\boldsymbol{\beta}))_{k,j} = \mathrm{E}\left(\sum_{i=1}^{n} x_{ij} x_{ik} \left(\frac{d\mu_i}{d\eta_i} \right)^2 \frac{1}{\mathrm{Var}(Y_i)} \right) = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2. \tag{16}$$

Letting $\tilde{\boldsymbol{\beta}}^k = (\tilde{\beta}_1^k, ..., \tilde{\beta}_p^k)^T$ be the vector of estimates at the kth iteration of the algorithm, the iteration of the scoring method to approximate the maximum likelihood estimates of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}}^{k+1} = \tilde{\boldsymbol{\beta}}^k + \left(J(\tilde{\boldsymbol{\beta}}^k) \right)^{-1} U(\tilde{\boldsymbol{\beta}}^k;\boldsymbol{y}), \tag{17}$$

where, as the notation indicates, the expected information matrix and the score function vector are evaluated at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}^k$. Note that, for certain problems, the inverse of $J(\boldsymbol{\beta})$ might not exist, in which case, generalized inverses can be employed.

To illustrate the equivalence of the scoring method with an IWLS method, using (17) we obtain

$$J(\tilde{\boldsymbol{\beta}}^k)\tilde{\boldsymbol{\beta}}^{k+1} = J(\tilde{\boldsymbol{\beta}}^k)\tilde{\boldsymbol{\beta}}^k + U(\tilde{\boldsymbol{\beta}}^k;\boldsymbol{y}). \tag{18}$$

Moreover, using (13) and (16), we can write

$$\begin{aligned} J(\boldsymbol{\beta}) &= \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{X} \\ J(\boldsymbol{\beta})\boldsymbol{\beta} + U(\boldsymbol{\beta};\boldsymbol{y}) &= \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{z}(\boldsymbol{\beta}), \end{aligned} \tag{19}$$

where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{W}(\boldsymbol{\beta})$ is an $n \times n$ diagonal matrix with elements

$$w_{ii}(\boldsymbol{\beta}) = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2, \quad i = 1, ..., n$$

and $\boldsymbol{z}(\boldsymbol{\beta})$ is an n-dimensional vector with elements

$$z_i(\boldsymbol{\beta}) = (y_i - \mu_i) \frac{d\eta_i}{d\mu_i} + \sum_{\ell=1}^{p} x_{i\ell}\beta_\ell, \quad i = 1, ..., n.$$

Hence, using (19), expression (18), that defines the iteration for the scoring method, can be written as

$$\left(\boldsymbol{X}^T \boldsymbol{W}(\tilde{\boldsymbol{\beta}}^k) \boldsymbol{X} \right) \tilde{\boldsymbol{\beta}}^{k+1} = \boldsymbol{X}^T \boldsymbol{W}(\tilde{\boldsymbol{\beta}}^k) \boldsymbol{z}(\tilde{\boldsymbol{\beta}}^k). \tag{20}$$

Expression (20) defines iterations for an IWLS algorithm with weights w_{ii} and (transformed) responses z_i, $i = 1, ..., n$. The approach is iterative, since, in general, both the w_{ii} and the z_i depend on $\boldsymbol{\beta}$.

References

Charnes, A., Frome, E.L., and Yu, P.L. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association*, 71, 169-171.

Draper, N.R., and Smith, H. (1981). *Applied Regression Analysis* (Second Edition). New York: John Wiley & Sons.

Efron, B., and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65, 457-487.

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

Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135, 370-384.

Appendix: Expectations for derivatives of log-likelihoods

In order to obtain the mean and variance of a random variable with a distribution that is a member of the exponential dispersion family (or the one-parameter exponential family), we need the following results:

$$\begin{aligned} \mathrm{E}\left(\frac{\partial \ell}{\partial \theta}\right) &= 0 \qquad (21) \\ \mathrm{E}\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) + \mathrm{E}\left(\left(\frac{\partial \ell}{\partial \theta}\right)^2\right) &= 0. \qquad (22) \end{aligned}$$

Here $\ell \equiv \ell(\theta; y) = \log f(y \mid \theta)$ is the log-likelihood of θ based on one realization y from some distribution (not necessarily from the exponential family) with density/mass function $f(y \mid \theta)$. For simplicity, we assume that θ is one dimensional taking values in $\mathbb{R}$ (or some subset of $\mathbb{R}$).

Establishing (21) and (22) requires certain regularity conditions for $f(y \mid \theta)$. These are basically conditions that allow reversing the order of integration and differentiation as needed below in (23) and (24) (see, e.g., Casella and Berger (1990), *Statistical Inference*, pp. 68-76). For example, distributions with support that depends on the parameter(s) create difficulties with regard to these conditions. However, they are satisfied by the exponential dispersion family.

Assume, without loss of generality, that Y is continuous. (We only need to replace integrals with sums in the expressions below if Y is discrete.) To verify (21), differentiate, with respect to θ, both sides of $\int f(y \mid \theta) dy = 1$ to obtain

$$\begin{aligned} \frac{\partial}{\partial \theta} \int f(y \mid \theta) dy &= 0 \\ \Rightarrow \int \frac{\partial f(y \mid \theta)}{\partial \theta} dy &= 0 \qquad (23) \\ \Rightarrow \int \frac{\partial \ell}{\partial \theta} f(y \mid \theta) dy &= 0 \\ \Rightarrow \mathrm{E}\left(\frac{\partial \ell}{\partial \theta}\right) &= 0. \end{aligned}$$

For (22), note that

$$\frac{\partial^2 \ell}{\partial \theta^2} = \frac{\partial}{\partial \theta}((f(y \mid \theta))^{-1} \frac{\partial f(y \mid \theta)}{\partial \theta}) = ((f(y \mid \theta))^{-1} \frac{\partial^2 f(y \mid \theta)}{\partial \theta^2}) - ((f(y \mid \theta))^{-1} \frac{\partial f(y \mid \theta)}{\partial \theta})^2.$$

Hence

$$\begin{aligned} \mathrm{E}\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) + \mathrm{E}\left((\frac{\partial \ell}{\partial \theta})^2\right) &= \mathrm{E}\left((f(y \mid \theta))^{-1} \frac{\partial^2 f(y \mid \theta)}{\partial \theta^2}\right) \\ &= \int \frac{\partial^2 f(y \mid \theta)}{\partial \theta^2} dy \\ &= \frac{\partial^2}{\partial \theta^2} \int f(y \mid \theta) dy \qquad (24) \\ &= 0. \end{aligned}$$