Generalized linear models

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 5, Hilary Term 2009

February 4, 2009

A generalized linear model is based on a family the form

$$f(y;\theta,\phi) = b(y,\phi)e^{\{y\theta - c(\theta)\}/d(\phi)}. \tag{1}$$

For ϕ fixed and θ varying over all possible values, this is a one-dimensional exponential family with canonical statistic $t(y) = y$, canonical parameter $\theta^* = \theta/d(\phi)$, and cumulant generating function

$$\kappa\{\theta^*\} = \kappa\{\theta/d(\phi)\} = c(\theta)/d(\phi) = \log \int b(y,\phi)e^{y\theta^*}\,dy, \tag{2}$$

so

$$\mathbf{E}(Y) = \frac{\partial}{\partial\theta^*}\kappa\{\theta/d(\phi)\} = d(\phi)\frac{\partial}{\partial\theta}\kappa\{\theta/d(\phi)\} = c'(\theta)$$

and

$$\mathbf{V}(Y) = \frac{\partial^2}{\partial{\theta^*}^2}\kappa\{\theta/d(\phi)\} = d(\phi)^2\frac{\partial^2}{\partial\theta^2}\kappa\{\theta/d(\phi)\} = c''(\theta)d(\phi).$$

An exponential families with the canonical statistic $t(y) = y$ is also known as a *natural exponential family* (NEF), but terminology varies among authors so beware. Clearly, one can either consider the family (1) as an exponential family with canonical statistic $t(y) = y$ and canonical parameter $\theta^* = \theta/d(\phi)$, or let $t^*(y) = y/d(\phi)$ with parameter θ.

For varying ϕ, the situation is generally much more complex. Sometimes it is an exponential family, sometimes not. Sometimes it is not possible to have $d(\phi)$ varying independently of θ at all, e.g. in the Poisson case.

When $d(\phi)$ is varying, it is a strong restriction on the function $b(y, \phi)$ to assume that the cumulant generating function (2) has the form $\kappa(\theta/\phi) = c(\theta)/d(\phi)$. Indeed, models where this is true are known as *dispersion models*.

Since $\mathbf{V}(Y) = d(\phi)c''(\theta)$ we have $c''(\theta) > 0$ and hence the function $c'(\theta)$ is strictly increasing in θ. We can therefore parametrize the family with its mean μ and define $\theta(\mu)$ by the relation

$$\mu = \mathbf{E}(Y) = c'(\theta), \quad \theta(\mu) = c'^{-1}(\mu)$$

and define the *variance function*

$$v(\mu) = c''\{\theta(\mu)\}$$

so now

$$\mathbf{V}(Y) = d(\phi)v(\mu)$$

and we can readily think of ϕ as a *dispersion parameter.*

An important fact is that *the variance functions identifies the family* in the sense that two families of densities which both have the form (1) and have the same variance function $v(\mu)$, must be identical.

Common variance functions for standard families are

Normal	Poisson	Binomial	Gamma	Inverse Gaussian
1	μ	$\mu(1-\mu)$	μ^2	μ^3

Not all functions $v(\mu)$ can occur as variance functions.

For example, *a function of the form $v(\mu) = \mu^\alpha$ is a variance function for a NEF if $\alpha \leq 0$ or $1 \leq \alpha < \infty$, but not if $0 < \alpha < 1$*.

Generalized linear models describe independent samples of the form $Y = (Y_1, \ldots, Y_n)$ where each Y_i is a one-dimensional *response* to *covariates* $x_i = (x_{i1}, \ldots, x_{ip})$ having distribution of the form (1), with expectations μ_i and dispersions $d_i(\phi)$.

For simplicity we assume $d_i(\phi) = \phi$ although $d_i(\phi) = \phi/w_i$ with w_i being a known *weight* may be appropriate in some cases. *Formally we assume ϕ known for the moment.*

The *saturated model* makes no further restriction on the parameters μ_i and the maximum likelihood estimator under this model is therefore given as

$$\hat{\mu} = Y,$$

provided the base exponential family is regular.

More generally, we restrict the vector of expectations $\mu = (\mu_1, \ldots, \mu_n)^\top$ through a *linear predictor* $\eta_i = x_i\beta$ written in matrix form as

$$\eta = X\beta$$

where x_i are the rows of X and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown parameters, and a *link function* g relating the linear predictor to the mean as

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

Here care should be taken in the choice of link function, as the parameter space for β must be restricted so that this equation makes sense.

A special role is played by the link function $g(\mu) = \theta(\mu) = c'^{-1}(\mu)$ which is known as *canonical link.*

If we consider the likelihood function we get

$$l(\beta) = \log L(\beta) = \sum_i \{y_i\theta_i - c(\theta_i)\}/\phi = \{y^\top\theta - \sum_i c(\theta_i)\}/\phi,$$

where now

$$\theta_i = \theta(\mu_i) = \theta\{g^{-1}(\eta_i)\} = \theta\{g^{-1}(x_i\beta)\}.$$

If g is the canonical link function, we have $g(\mu) = \theta(\mu)$ and hence $\theta_i = x_i\beta$. This then yields

$$l(\beta) = \{y^\top X\beta - \sum_i c(x_i\beta)\}/\phi = \{(X^\top y)^\top\beta - \sum_i c(x_i\beta)\}/\phi$$

and hence the family of joint distributions is a linear and canonical exponential family with canonical statistic $t(y) = X^\top y$ and β/ϕ as the canonical parameter.

Thus, the likelihood equation for a fixed ϕ again equates the expectation of the sufficient statistic to the observed value. Interpreting vector functions componentwise this has the simple form

$$X^\top \mu(\beta) = X^\top y$$

or equivalently

$$X^\top \{y - \mu(\beta)\} = 0$$

expressing that the residual $y - \mu(\beta)$ is orthogonal to all columns of X.

From general theory of exponential families it is known that *there is at most one solution $\hat{\beta}$ to this equation,* despite the fact that the equation typically is non-linear in β, as $\mu(\beta) = g^{-1}(X\beta)$.

For a general link function, the score statistic can be written in the form

$$S(\beta) = Z^\top W\{y - \mu(\beta)\}/\phi \tag{3}$$

where Z is a matrix with elements

$$Z(\beta)_{ij} = \frac{\partial \eta_i}{\partial \beta_j}$$

and $W(\beta)$ is a diagonal matrix with diagonal elements equal to $W_{ii} = 1/v\{\mu_i(\beta)\}$.

Now, in contrast to the case of a canonical link function, *the corresponding likelihood equation may have multiple solutions or none at all* and the linearity of the predictor can not be used to resolve this fact.

Fisher's method of scoring leads to a *iterative weighted least squares regression* procedure (IRLS) for solving the likelihood equations.

This fact can now be used for all generalized linear models simultaneously, only the calculation of the matrix Z and the weights W being special to the model considered, depending in a simple way on the link and variance functions. Details are omitted here, but much of the success of these models hinges on this computational fact.

The goodness of fit of a specific generalized linear model is assessed in the usual way using the *deviance*

$$\begin{aligned} D(\hat{\mu}; y) &= -2\{l(\hat{\mu}; y) - l(y; y)\} \\ &= -2\{l_1(\hat{\mu}; y) - l_1(y; y)\}/\phi = D_1(\hat{\mu}; y)/\phi, \end{aligned}$$

where $l(y; y)$ is the maximized log-likelihood in the saturated model and $l(\hat{\mu}; y)$ is the maximized log-likelihood in the model considered.

The symbol l_1 is used for the log-likelihood in the case $\phi = 1$ and similarly for D_1.

Under reasonable assumption on the behaviour of the covariates x_i, D can be shown to be asymptotically distributed as a χ^2-distribution with degrees of freedom $n - p$ where X is assumed to have full rank p.

In the situation, where the dispersion parameter ϕ is considered *unknown* it is therefore customary to use the estimator

$$\tilde{\phi} = \frac{D_1(\hat{\mu}, Y)}{n - p}.$$

Note that this is *not* a maximum likelihood estimator, and there are good reasons for not using the MLE:

Firstly, the problem of finding the MLE of ϕ could be computationally very difficult in general, and the computational problem very different for different variance functions.

Secondly, there would be a problem with the *nuisance parameter* β distorting the estimate, in particular if the dimension p of β is large.

The estimate for ϕ used is thus based on *'approximate marginal likelihood'*, estimating ϕ on the basis of the approximate χ^2-distribution for the deviance. The MLE of μ is the same for all values of ϕ and is therefore appropriate as is.