Outline of GLMs

Definitions

This is a short outline of GLM details, adapted from the book "Nonparametric Regression and Generalized Linear Models", by Green and Silverman.

The responses Y_i have density (or mass function)

$$p(y_i; \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right) \tag{1}$$

which becomes a one parameter exponential family in θ_i if the scale parameter ϕ is known. This is simpler than McCullagh and Nelder's version: they use $a(\phi)$ or even $a_i(\phi)$ in place of ϕ. The $a_i(\phi)$ version is useful for data whose precision varies in a known way, as for example when Y_i is an average of n_i iid random variables.

There are two simple examples to keep in mind. Linear regression has $Y_i \sim N(\theta_i, \sigma^2)$ where $\theta_i = x_i\beta$. Logistic regression has $Y_i \sim Bin(m, \theta_i)$ where $\theta_i = (1 + \exp(-x_i\beta))^{-1}$. It is a good exercise to write these models out as in (1), and identify ϕ, b and c. Consider also what happens for the models $Y_i \sim N(\theta_i, \sigma_i^2)$ and $Y_i \sim Bin(m_i, \theta_i)$.

The mean and variance of Y_i can be determined from its distribution given in equation (1). They must therefore be expressable in terms of ϕ, θ_i, and the functions $b(\cdot)$ and $c(\cdot, \cdot)$. In fact,

$$E(Y_i; \theta_i, \phi) = \mu_i = b'(\theta_i) \tag{2}$$

with b' denoting a first derivative, and

$$V(Y_i; \theta_i, \phi) = b''(\theta_i)\phi. \tag{3}$$

Equations (2) and (3) can be derived by solving

$$E(\partial \log p(y_i; \theta_i, \phi)/\partial\theta_i) = 0$$

and

$$E(-\partial^2 \log p(y_i; \theta_i, \phi)/\partial\theta_i^2) = E((\partial \log p(y_i; \theta_i, \phi)/\partial\theta_i)^2).$$

To generalize linear models, one links the parameter θ_i to (a column of) predictors x_i by

$$G(\mu_i) = x_i'\beta \tag{4}$$

for a “link function” G. Thus

$$\theta_i = (b')^{-1}(G^{-1}(x_i'\beta)). \tag{5}$$

The choice $G(\cdot) = b'^{-1}(\cdot)$ is convenient because it makes $\theta_i = x_i\beta$, with the result that $X'Y = \sum_i x_i'y_i$ becomes a sufficient statistic for β (given ϕ). Of course real data are under no obligation to follow a distribution with this canonical link. But if they are close to this link then sufficient statistics allow one a great reduction data in set size.

Estimation

The log likelihood function is

$$l(\theta, \phi) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right) \tag{6}$$

where θ subsumes all of the θ_i. It could also be written as a function of β and ϕ because (given the x_i), β determines all the θ_i.

The main way of estimating β is by maximizing (6). The fact that $G(\mu_i) = x_i\beta$ suggests a crude approximate estimate: regress $G(y_i)$ on x_i, perhaps modifying y_i in order to avoid violating range restrictions (such as taking $\log(0)$), and accounting for the differing variances of the observations.

Fisher scoring is the widely use technique for maximizing the GLM likelihood over β. The basic step is

$$\beta^{(k+1)} = \beta^{(k)} - \left(E\left(\frac{\partial^2 l}{\partial\beta\partial\beta'} \right) \right)^{-1} \frac{\partial l}{\partial\beta} \tag{7}$$

where expectations are taken with $\beta = \beta^{(k)}$. This is like a Newton step, except that the Hessian of l is replaced by it’s expectation. The expectation takes a simpler form. This is similar to (generalizes?) what happens in the Gauss-Newton algorithm, where individual observation Hessians multiplied by residuals are dropped. (It clearly would not work to replace $\partial l/\beta$ by its expectation!)

Fisher scoring simplifies to

$$\beta^{(k+1)} = (X'WX)^{-1}X'Wz \tag{8}$$

where W is a diagonal matrix with

$$W_{ii} = (G'(\mu_i)^2 b''(\theta_i))^{-1}. \tag{9}$$

and

$$z_i = (Y_i - \mu_i)G'(\mu_i) + x_i\beta. \tag{10}$$

Both equations (9) and (10) use $\beta^{(k)}$ and the derived values of $\theta_i^{(k)}$ and $\mu_i^{(k)}$.

The iteration (8) is known as "iteratively reweighted least squares", or IRLS. The weights W_{ii} have the usual interpretation as reciprocal variances: $b''(\theta_i)$ is proportional to the variance of Y_i and the $G'(\mu_i)$ factor in z_i is squared in W_{ii}. The constant ϕ does not appear. More precisely: it appeared in (7) twice and cancelled out.

Fisher scoring may also be written

$$\beta^{(k+1)} = \beta^{(k)} + (X'WX)^{-1}X'Wz^* \tag{11}$$

where $z_i^* = (Y_i - \mu_i)G'(\mu_i)$.

Inference

The saturated model S is one with n parameters $\widetilde{\theta}_i$, one for each y_i. The value of θ_i in the saturated model is the maximizer of (1), commonly y_i itself. In the $N(\theta_i, \sigma^2)$ model this gives a zero sum of squares for the saturated model. [Exercise: what is the likelihood for a saturated normal model?] Note that if $x_i = x_j$, so that for any β $x_i\beta = x_j\beta$, the saturated model can still have $\widetilde{\theta}_i \neq \widetilde{\theta}_j$.

Let $l_{\max}$ be the log likelihood of S. In the widely used models this is finite. (In some cases saturating a model can lead to an infinite likelihood. Consider the normal distribution where one picks a mean *and* a variance for each observation.)

The scaled deviance is

$$D^* = 2(l_{\max} - l(\theta(\hat{\beta}))) \tag{12}$$

and the unscaled deviance is

$$D = \phi D^* = 2\sum_{i=1}^{n}\Big(Y_i(\widetilde{\theta}_i - \hat{\theta}_i) - b(\widetilde{\theta}_i) + b(\hat{\theta}_i)\Big) = \sum_{i=1}^{n} d_i. \tag{13}$$

The scaled deviance D^* (and the unscaled when $\phi = 1$) is asymptotically χ^2_{n-p} where p is the number of parameters in β, under some restrictive conditions. The usual likelihood ratio asymptotics do not apply because the degrees of freedom $n - p$ is growing as fast as the number of observations n. In a limit with each observation becoming "more normal" and the number of observations remaining fixed a chisquare result can obtain. (Some examples: $\text{Bin}(m_i, p_i)$ observations with increasing m_i and $m_i p_i(1-p_i)$ bounded away from zero, and $\text{Poi}(\lambda_i)$ observations with λ_i increasing.) There are however cases where a chisquare limit holds with n increasing (see Rice's introductory statistics book).

When the χ^2_{n-p} approximation is accurate we can use it to test goodness of fit. A model with too large a deviance doesn't fit the data well.

Differences between scaled deviances of nested models have a more nearly chisquare distribution than the deviances themselves. Because the degrees of freedom between two models is fixed these differences are also chisquared in the usual asymptotic setup, of increasing n.

The deviance is additive so that for models $M_1 \subseteq M_2 \subseteq M_3$, the chisquare statistic for testing M_1 within M_3 is the sum of those for testing M_1 within M_2 and M_2 within M_3.

The deviance generalizes the sum of squared errors (and D^* generalizes the sum of squares normalized by σ^2). Another generalization of sum of squared errors is Pearson's chisquare

$$\chi^2 = \phi \sum_{i=1}^{n} \frac{(Y_i - E(Y_i))^2}{V(Y_i)} = \sum_{i=1}^{n} \frac{(Y_i - b'(\theta_i))^2}{b''(\theta_i)}, \tag{14}$$

often denoted by X^2 to indicate that it is a sample quantity. Pearson's chisquare can also be used to test nested models, but it does not have the additivity that the deviance does. It may also be asymptotically chisquared, but just as for the deviance, there are problems with large "sparse" models. (Lots of observations but small m_i or λ_i.) Pearson's chisquare is often used to estimate ϕ by

$$\hat{\phi} = \frac{\chi^2}{n-p}. \tag{15}$$

Residuals can be defined by partitioning either the deviance or Pearson's χ^2. The deviance residuals are

$$r_i^{(D)} = \text{sign}(Y_i - \mu_i)\sqrt{d_i} \tag{16}$$

where d_i is from (13) and the Pearson residuals are

$$r_i^{(P)} = \frac{Y_i - \mu_i}{\sqrt{b''(\theta_i)}}. \tag{17}$$

These can also be "leverage adjusted" to reflect the effect of x_i on the variance of the different observations. See references in Green and Silverman.

These residuals can be used in plots to test for lack of fit of a model, or to identify which points might be outliers. They need not have a normal distribution, so qq plots of them can't necessarily diagnose a lack of fit.

Approximate confidence intervals may be obtained by profiling the likelihood function and referring the difference in (scaled) deviance values to calibrate the coverage levels. Computationally simpler confidence statements can be derived from the asymptotic covariance matrix of $\hat{\beta}$

$$-\left(E\frac{\partial^2 l}{\partial\beta\partial\beta'}\right)^{-1} = \phi(X'WX)^{-1} \tag{18}$$

(I think the formula that Green and Silverman give (page 97) instead of (18) is wrong.....they have partials with respect to a vector η with components $\eta_i = x_i\beta$.) These simple confidence statements are much like the linearization inferences in nonlinear least squares. They judge $H_0 : \beta - \beta_0$ not by the likelihood at β_0 but by a prediction of what that likelihood would be based on an expansion around $\beta - \hat{\beta}$.

Overdispersion

Though our model may imply that $V(Y_i) = \phi b''(\theta_i)$ in practice it can turn out that the model is too simple. For example a model in which Y_i is a count may assume it to be a sum of independent identically distributed Bernoulli random variables. But non-constant success probability or dependence among the Bernoulli variables can cause $V(Y_i) \neq m_i p_i(1 - p_i)$. We write this as

$$V(Y_i) = b''(\theta_i)\phi\sigma_i^2 \tag{19}$$

for some overdispersion parameter $\sigma_i^2 \neq 1$.

One can model the overdispersion in terms of x_i and possibly other parameters as well. But commonly a very simple model with a constant value $\sigma_i^2 = \sigma^2$ is used. Given that variances are usually harder to estimate than means, it is reasonable to use a smaller model for the overdispersion than for θ_i.

A constant value of σ^2 is equivalent to replacing ϕ by $\phi\sigma^2$, so an estimated quantity $\phi\hat{\sigma}^2$ can be simplified to $\hat{\phi}$.

Given an estimate of the amount of overdispersion one can use it to adjust the estimated covariance matrix of $\hat{\beta}$: multiply the naive estimate by $\hat{\sigma}^2$. Similarly one can adjust the chisquare statistics: divide the naive ones by $\hat{\sigma}^2$.

Quasi and empirical likelihoods

The asymptotic good behavior of generalized linear modelling depends on getting the relationship between $\mu = E(Y) = b'(\theta)$ and $V(Y) = \phi b''(\theta)$ right. The method of quasi likelihood involves simply specifying how $V(Y)$ depends on μ, without specifying a complete likelihood function. Asymptotically the inferences are justified for any likelihood having the postulated mean-variance relationship. McCullagh and Nelder (Chapter 9) discuss this. In quasi-likelihood, the derivative of the log likelihood is replaced by $(Y - \mu)/\sigma^2 V(\mu)$ for the postulated function $V(\mu)$.

The method of empirical likelihood can also be used with generalized linear models. It can give asymptotically justified confidence statements and tests, and does not even require a correctly specified relationship between $E(Y)$ and $V(Y)$. This was studied by Eric Kolaczyk in Statistica Sinica in 1994.

Examples of GLMs

McCullagh and Nelder give a lot of examples of uses of GLMs. They consider models for cell counts in large tables, models for survival times and models for variances.