Generalized Linear Models

Gene Katsevich, Stephen Bates

July 14, 2019

1 Definition

Let

$$f_\eta(y) = \exp(\eta y - \psi(\eta)) f_0(y)$$

be a one-parameter exponential family with natural parameter η, cumulant generating function $\psi(\eta)$, and base measure $f_0(y)$. Let

$$\mu = \mathbb{E}_\eta[y] = \dot{\psi}(\eta), \quad W(\eta) = \mathrm{Var}_\eta[y] = \ddot{\psi}(\eta).$$

We can parameterize the distribution using either the natural parameter η or the mean parameter μ. Letting $g = \dot{\psi}^{-1}$, we have

$$\eta = g(\mu).$$

Now, suppose that we have data $(x_i, y_i) \in \mathbb{R}^{p+1}$. A canonical generalized linear model for this data is

$$y_i \sim f_{\eta_i}; \quad \eta_i = x_i^T \beta, \tag{1}$$

for some $\beta \in \mathbb{R}^p$. Hence, we model the natural parameter as a linear function of our covariates, and the result is a new exponential family with probability density

$$p(y) = \prod_{i=1}^n e^{\beta^T x_i y_i - \psi(x_i^T \beta)} f_0(y) \tag{2}$$

which is now an exponential family with parameter $\beta \in \mathbb{R}^p$.

2 Examples

Linear regression. Consider a linear regression model with known variance. After rescaling, we can assume $\sigma^2 = 1$, so we are working with the exponential family

$$f_\mu(y) = \exp\left(y\mu - \frac{1}{2}\mu^2\right) \exp\left(-\frac{1}{2}y^2\right); \quad \eta = \mu.$$

In this case the mean parameter and natural parameter coincide. We have

$$y_i \sim N(\mu_i, 1), \quad \eta_i = \mu_i = x_i^T \beta.$$

Logistic regression. Suppose we have binary responses $y_i \in \{0, 1\}$. Then, it is appropriate to work with the Bernoulli family:

$$f_\mu(y) = \exp\left(y \log \frac{\mu}{1-\mu} + \log(1-\mu)\right), \quad \eta = \log \frac{\mu}{1-\mu}.$$

We have

$$y_i \sim \mathrm{Ber}(\mu_i), \quad \eta_i = \mathrm{logit}(\mu_i) = x_i^T \beta.$$

Poisson regression. Suppose we have count responses $y_i \in \mathbb{N}$. Then, we can work with the Poisson family:

$$f_\mu(y) = \exp(y \log \mu - \mu)\frac{1}{y!}, \quad \eta = \log \mu.$$

We have

$$y_i \sim \mathrm{Poi}(\mu_i), \quad \eta_i = \log(\mu_i) = x_i^T \beta.$$

3 Parameter estimation

Given data (x_i, y_i), $i = 1, \ldots, n$, we can fit the parameters β by maximum likelihood. Let $X \in \mathbb{R}^{n \times p}$ be the design matrix, whose ith row is x_i^T, and let $y \in \mathbb{R}^n$ be the response vector. For a general mean vector $\mu \in \mathbb{R}^n$, the log-likelihood of the data is

$$\ell(\eta; y) = \sum_{i=1}^n \left(\eta_i y_i - \psi(\eta_i)\right).$$

From (1), this is equivalent to an exponential family with parameter β :

$$\ell(\beta; y) = \sum_{i=1}^n \left(x_i^T \beta y_i - \psi(x_i^T \beta)\right) = \beta^T X^T y - \sum_{i=1}^n \psi(x_i^T \beta). \tag{3}$$

Differentiating in β, we get the score equation:

$$0 = \dot{\ell}(\hat{\beta}; y) = X^T y - \sum_{i=1}^n x_i \dot{\psi}(x_i^T \hat{\beta}) = X^T y - \sum_{i=1}^n x_i \hat{\mu}_i = X^T y - X^T \hat{\mu}.$$

Hence, the score equation states that the observed value of the sufficient statistic $X^T y$ must equal its expected value under $\hat{\beta}$, which is $X^T \hat{\mu}$. Thus, we need to solve

$$X^T (y - \hat{\mu}) = 0. \tag{4}$$

Note that in the linear model, $\hat{\mu} = X\hat{\beta}$, so the score equation (4) reduces to the normal equations.

Unlike the special case of the linear model, in general $\hat{\mu}$ is a nonlinear function of $\hat{\beta}$, and so (4) cannot be solved explicitly. Notice that (3) is concave, however, since properties of exponential families imply that ϕ is convex. As a result, maximizing the likelihood can reliably be done numerically.

Newton-Raphson algorithm for the MLE

The standard algorithm for finding a solution to (4) is the Newton-Raphson algorithm carried out on $\dot{\ell}$. In particular, the second derivative to ℓ is given by

$$\ddot{\ell}(\beta) = -X^T W_\beta X,$$

where W_β is a $n \times n$ diagonal matrix with entries $\ddot{\psi}(x_i^T\beta)$. One step of the Newton-Raphson algorithm is obtained by setting the taylor approximation to (4), taken at the previous value of $\hat{\beta}$ to be zero:

$$\begin{aligned} 0 &= X^T(y - \mu_{\hat{\beta}^{(t)}}) - X^T W_{\hat{\beta}^{(t)}} X(\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}) \\ \hat{\beta}^{(t+1)} &= \hat{\beta}^{(t)} + (X^T W_{\hat{\beta}^{(t)}} X)^{-1} X^T(y - \mu_{\hat{\beta}^{(t)}}). \end{aligned}$$

This update can be rewritten as

$$\hat{\beta}^{(t+1)} = (X^T W_{\hat{\beta}^{(t)}} X)^{-1} X^T W_{\hat{\beta}^{(t)}} \left(W_{\hat{\beta}^{(t)}}^{-1}(y - \mu_{\hat{\beta}^{(t)}}) + X\hat{\beta}^{(t)} \right),$$

which has a nice interpretation as the solution to a weighted least squares problem, as discussed next.

Interpretation as a Gaussian approximation

For GLMs, the above algorithm is sometimes called the *iteratively reweighted least squares* algorithm because it can be formulated as a sequence of least-squares problems. The idea is the following. Suppose we have a current guess $\hat{\beta}^{(t)}$ of the parameters. Then, we approximate (1) using a model with heteroskedastic normal errors:

$$y_i = \mu_i(\beta) + \epsilon_i, \quad \epsilon_i \sim N(0, W_{\beta^{(t)}(ii)}), \tag{5}$$

since $W_{\beta^{(t)}(ii)}$ is the variance of y_i when the true parameter is $\hat{\beta}^{(t)}$. Next, we use a Taylor series approximation of $\mu(\beta)$ to get

$$\mu(\beta) = \hat{\mu}^{(t)} + \hat{W}^{(t)}(X\beta - X\hat{\beta}^{(t)}), \tag{6}$$

where $\hat{W}^{(t)} = W_{\hat{\beta}^{(t)}}$. Putting together (5) and (6), we get the approximate weighted linear model

$$y = \hat{\mu}^{(t)} + \hat{W}^{(t)}(X\beta - X\hat{\beta}^{(t)}) + \epsilon, \quad \epsilon \sim N(0, \hat{W}^{(t)}),$$

which we can rearrange to obtain

$$\hat{z}^{(t)} = X\beta + \epsilon, \quad \epsilon \sim N(0, (\hat{W}^{(t)})^{-1}), \tag{7}$$

where

$$\hat{z}^{(t)} = X\hat{\beta}^{(t)} + (\hat{W}^{(t)})^{-1}(y - \hat{\mu}^{(t)}).$$

is the *adjusted response variable.* The solution to this weighted linear model is

$$\hat{\beta}^{(t+1)} = (X^T\hat{W}^{(t)}X)^{-1}X^T\hat{W}^{(t)}\hat{z}^{(t)}. \tag{8}$$

The algorithm proceeds by iteratively calculating the linear approximation (7) and then solving it via (8) to update the parameter estimates. After rearranging terms, this can be seen to be equivalent to the Newton-Raphson method above.

4 Inference

Now that we can fit generalized linear models, how does inference work? We might be interested in finding standard errors and confidence intervals for $\hat{\beta}$, or we might want to test hypotheses with respect to these parameters.

4.1 Confidence intervals

Using standard likelihood theory, we get the following asymptotic distribution of $\hat{\beta}$:

$$\hat{\beta} \sim \mathcal{N}(\beta, (X^T W X)^{-1}).$$

Note that this is the same expression we would get from the approximate weighted linear model (7). Importantly, note that W itself depends on β. As usual, we can form standard errors and confidence intervals for $\hat{\beta}$ based on this asymptotic distribution.

4.2 Hypothesis testing

Suppose we have two nested models $M_1 \subset M_2$, and we want to test the null hypothesis that model M_1 is adequate. In linear regression, we would usually use the F test, the numerator of which is $\text{RSS}(M_1) - \text{RSS}(M_2)$; i.e. the amount by which switching from M_1 to M_2 decreases the RSS.

For generalized linear models, the RSS is no longer appropriate, so instead we use the *deviance*. The deviance of a model M is defined as twice the increase in log-likelihood you get by switching from model M to the *saturated model*. The saturated model is where we fit $y_i \sim f_{\eta_i}$, letting η_i be unrestricted (instead of being parameterized by covariates). From exponential family theory, the MLE in the unrestricted case is $\hat{\mu}_i = y_i$. Hence, the definition of the deviance is

$$D(y; \mu(\hat{\beta}_M)) = 2(\ell(y; y) - \ell(y; \mu(\hat{\beta}_M))).$$

For the linear model, note that

$$D(y; \hat{\mu}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2,$$

so indeed it coincides with the RSS.

The deviance itself can be viewed as a statistic for a goodness of fit test for model M, and under the null hypothesis that model M fits, it has an asymptotic distribution of

$$D(y; \mu(\hat{\beta}_M)) \sim \chi^2_{n-|M|},$$

where $|M|$ is the number of predictors in $|M|$. More generally, to test model M_1 versus model M_2, we would use the difference in deviances $D(y; \mu(\hat{\beta}_{M_1})) - D(y; \mu(\hat{\beta}_{M_2}))$, which is twice the usual likelihood ratio test statistic. Under the null hypothesis that the smaller model M_1 fits, it has the asymptotic distribution

$$D(y; \mu(\hat{\beta}_{M_1})) - D(y; \mu(\hat{\beta}_{M_2})) \sim \chi^2_{|M_2|-|M_1|}.$$