Generalized Linear Models

Simon Jackman
Stanford University

Generalized linear models (GLMs) are a large class of statistical models for relating responses to linear combinations of predictor variables, including many commonly encountered types of dependent variables and error structures as special cases. In addition to regression models for continuous dependent variables, models for rates and proportions, binary, ordinal and multinomial variables and counts can be handled as GLMs. The GLM approach is attractive because it (1) provides a general theoretical framework for many commonly encountered statistical models; (2) simplifies the implementation of these different models in statistical software, since essentially the same algorithm can be used for estimation, inference and assessing model adequacy for all GLMs.

The canonical treatment of GLMs is McCullagh and Nelder (1989), and this review closely follows their notation and approach. Begin by considering the familiar linear regression model, $y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$, where $i = 1, \ldots, n$, y_i is a dependent variable, $\mathbf{x}_i$ is a vector of k independent variables or predictors, $\boldsymbol{\beta}$ is a k-by-1 vector of unknown parameters and the ε_i are zero-mean stochastic disturbances. Typically, the ε_i are assumed to be independent across observations with constant variance σ^2, and distributed normal. That is, the normal linear regression model is characterized by the following features:

1. **stochastic component:** the y_i are usually assumed to have independent normal distributions with $E(y_i) = \mu_i$, with constant variance σ^2, or $y_i \overset{\text{iid}}{\sim} N(\mu_i, \sigma^2)$
2. **systematic component:** the covariates $\mathbf{x}_i$ combine linearly with the coefficients to form the linear predictor $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$.
3. **link between the random and systematic components:** the linear predictor $\mathbf{x}_i\boldsymbol{\beta} = \eta_i$ is a function of the mean parameter μ_i via a *link* function, $g(\mu_i)$. Note that for the normal linear model, g is an identity.

GLMs follow from two extensions of this setup: (1) stochastic components following distributions other than the normal; (2) link functions other than the identity.

Stochastic Assumptions. GLMs may be used to model variables following distributions in the exponential family: i.e.,

$$\begin{aligned} f(y;\theta,\psi) &= \exp\left\{\frac{y\theta - b(\theta)}{a(\psi)} + c(y;\psi)\right\}, \text{ or} \\ \log f(y;\theta,\psi) &= \frac{y\,\theta - b(\theta)}{a(\psi)} + c(y;\psi) \end{aligned} \tag{1}$$

where ψ is a dispersion parameter. For instance, the normal distribution

$$\begin{aligned} f(y;\theta,\psi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}, \text{ with log density} \\ \log f(y;\theta,\psi) &= \frac{y\,\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right) \end{aligned}$$

results with $\theta = \mu, \psi = \sigma^2, a(\psi) = \psi, b(\theta) = \theta^2/2$ and $c(y;\psi) = -\frac{1}{2}(y^2/\sigma^2 + \log(2\pi\sigma^2))$. Many other commonly used distributions are in the exponential family. In addition to the examples listed in Table 1, several other distributions are in the exponential family, including the beta, multinomial, Dirichlet, and Pareto. Distributions that are not in the exponential family but are used for statistical modeling include the student's *t* and uniform distributions.

Link Functions. In theory, link functions $\eta_i = g(\mu_i)$ can be any monotonic, differentiable function. In practice, only a small set of link functions are actually utilized. In particular, links are chosen such that the *inverse link*, $\mu_i = g^{-1}(\eta_i)$ is easily computed, and so that g^{-1} maps from $\mathbf{x}_i\boldsymbol{\beta} = \eta_i \in \mathbb{R}$ into the set of admissible values for μ_i. Table 1 lists the canonical links: for instance, a log link is usually used for the Poisson model, since while $\eta_i = \mathbf{x}_i\boldsymbol{\beta} \in \mathbb{R}$, because y_i is a count, we have $\mu_i \in 0, 1, \ldots$. For binomial data, the link function maps from $0 < \pi < 1$ to $\eta_i \in \mathbb{R}$, and three links are commonly used: (1) *logit*: $\eta_i = \log(\pi_i/(1-\pi_i))$; (2) *probit*: $\eta_i = \psi^{-1}(\mu_i)$, where $\psi(\cdot)$ is the normal cumulative distribution function; (3) *complementary log-log*: $\eta_i = \log(-\log(1-\mu_i))$. Note that binary data are handled in the GLM framework as special cases of binomial data.

Estimation and Inference for $\boldsymbol{\beta}$. Equation 1 provides an expression for the log density of the dependent variable for a GLM; summing over these log-densities provides an expression for the log-likelihood,

$$l(\boldsymbol{\beta}, \psi; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} \left[\frac{y_i\,\theta_i - b(\theta_i)}{a(\psi)} + c(y_i;\psi)\right]. \tag{2}$$

Maximum likelihood estimates of $\boldsymbol{\beta}$ are generally not available in closed form, but can be obtained via an algorithm known as iteratively weighted least squares (IWLS). IWLS is one of the key practical ways in which GLMs are in fact "general", providing MLEs for a wide class of commonly used models. IWLS underlies the implementation of GLMs in software

	Normal	Poisson	Binomial	Gamma
Notation	$N(\mu, \sigma^2)$	$P(\mu)$	$B(n, \pi)/n$	$G(\mu, \nu)$
log-density	$\frac{1}{\sigma^2}\left(y\mu - \frac{1}{2}\mu^2 - \frac{1}{2}y^2\right) - \frac{1}{2}\log(2\pi\sigma^2)$	$y\log\mu - \mu - \log(y!)$	$n\left[y\log(\frac{\pi}{1-\pi}) + \log(1-\pi)\right] + \log\binom{n}{ny}$	$\nu\left(-\frac{y}{\mu} - \log\mu\right) + \nu\log y + \nu\log\nu - \log\Gamma(\nu)$
Range of y	$(-\infty, \infty)$	$0, 1, \ldots, \infty$	$z/n,\ z \in \{0, 1, \ldots, n\}$	$(0, \infty)$
Dispersion parameter, ψ	σ^2	1	n^{-1}	ν^{-1}
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$-\log(-\theta)$
$c(y; \psi)$	$-\frac{1}{2}\left(\frac{y^2}{\psi} + \log(2\pi\psi)\right)$	$-\log y!$	$\log\binom{n}{ny}$	$\nu\log(\nu y) - \log y - \log\Gamma(\nu)$
$\mu(\theta) = E(y; \theta)$	θ	$\exp(\theta)$	$e^\theta/(1 + e^\theta)$	$-1/\theta$
Canonical link, $\theta(\mu)$	identity ($\theta = \mu$)	$\log(\mu)$	$\log\left(\frac{\pi}{1-\pi}\right)$	μ^{-1}
Variance function, $V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2
Deviance function, D_M	$\sum(y_i - \hat{\mu}_i)^2$	$2\sum\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i\right]$	$2\sum\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i-y_i}{n_i-\hat{\mu}_i}\right)\right]$	$2\sum\left[-\log\left(\frac{y_i}{\mu_i}\right) + \frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right]$

Table 1: Common Distributions in the Exponential Family used with GLMs.

packages such as S-Plus and R; that is, rather than coding up problem-specific routines for optimizing likelihood functions with respect to unknown parameters, the GLM framework and IWLS provides a common framework for implementing models in statistical software. Briefly, iteration t of IWLS consists of:

1. form the *working responses* $z_i^{(t)} = \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \left(\frac{d\eta}{d\mu}\right)_i^{(t)}$, where $\eta_i^{(t)} = \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(t\text{-}1)}$ and $\mu_i^{(t)} = g^{-1}(\eta_i^{(t)})$. For the standard models presented in Table 1, the derivative of the link $d\eta/d\mu$ is easy to compute.

2. form *working weights* $W_i^{(t)} = \left[\left(\frac{d\eta}{d\mu}\right)_i^2 V_i^{(t)}\right]^{-1}$ where $V_i^{(t)} = V(\mu_i^{(t)})$ is referred to as the *variance function*; McCullagh and Nelder (1989, 29) show that $\text{var}(y_i) = b''(\theta)a(\psi)$, with the first term the variance function, and the second term the dispersion parameter. Table 1 lists the variance functions for commonly used GLMs.

3. run the weighted regression of the $z_i^{(t)}$ on the covariates $\mathbf{x}_i$ with weights $W_i^{(t)}$; designate the coefficients from this weighted regression as $\hat{\boldsymbol{\beta}}^{(t)}$ and proceed to the next iteration.

This algorithm can be repeated until convergence in $\hat{\boldsymbol{\beta}}$ or the log-likelihood, and the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is given by $V(\hat{\boldsymbol{\beta}}) = \hat{\psi}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ where $\mathbf{W} = \text{diag}(W_1, \ldots, W_n)$ computed at the final iteration. Asymptotically-valid standard errors for the coefficients are obtained by taking the square root of the leading diagonal of $V(\hat{\boldsymbol{\beta}})$. Note also that the IWLS algorithm does not involve the dispersion parameter $\hat{\psi}$, which is usually either fixed *a priori* by the model (e.g., for binomial or Poisson models) or (as in the Gaussian case) estimated after computing residuals with $\hat{\boldsymbol{\beta}}$. Further details on the algorithm appear in McCullagh and Nelder (1989, 41-43); Nelder and Wedderburn (1972) introduced the term "generalized linear model" when discussing how the IWLS method could be used to obtains MLEs for exponential-family models.

Analysis of Deviance. In the GLM framework, it is customary to use a quantity known as *deviance* to formally assess model adequacy and to compare models. Deviance statistics are identical to those obtained using likelihood ratio test statistics. For a data set with n observations, assume the dispersion parameter is known or fixed at $\psi = 1$, and consider two extreme models: (1) a one parameter *null* model, setting $E(y_i) = \hat{\mu}, \forall\, i$; (2) a n parameter *saturated* model, denoted S, fitting a parameter for each data point, setting $E(y_i) = \hat{\mu}_i = y_i$. For a normal regression, model (1) is an "intercept-only", and model (2) results by fitting a dummy variable for every observation. Any interesting statistical model M lies somewhere in between, with $p < n$ parameters, but the two extreme cases provide benchmarks for assessing the performance of model M. In particular, if $\hat{\theta}_i^M = \theta^M(\hat{\mu}_i)$ are the predictions of model M, and $\hat{\theta}_i^S = \theta^S(y_i) = y_i$ are the predictions of model S, then the *deviance* of model M is

$$D_M = 2 \sum_{i=1}^{n} \left[\left\{ y_i \, \hat{\theta}^S(y_i) - b(\hat{\theta}_i^S) \right\} - \left\{ y_i \, \hat{\theta}_i^M - b(\hat{\theta}_i^M) \right\} \right] . \tag{3}$$

and when the dispersion ψ is estimated or known not to be 1, the *scaled deviance* is $D_M^* = D_M/\psi$. Several features of deviance should be noted:

1. Deviance is minimized when $M = S$; alternatively, the log-likelihood of S, $l^S = l(\mathbf{y};\mathbf{y})$ is (by definition) the highest log-likelihood attainable with data $\mathbf{y}$.
2. The scaled deviance of model M is the likelihood ratio test statistic for the test of null hypothesis that the $n - p$ parameter restrictions in M relative to S are true, i.e.,

$$D_M^* = 2[l(\mathbf{y};\mathbf{y}) - l(\hat{\boldsymbol{\theta}}^S;\mathbf{y})].$$

3. Asymptotically, $D_M^* \sim \chi^2_{n-p}$, facilitating hypothesis tests.
4. If model $M_0 \subset M$ (M_0 nests in M) with $q < p$ parameters, then if ψ is known, the difference in scaled deviances also provides a likelihood ratio statistic, for the null hypothesis that the $p - q$ restrictions in M_0 relative to M are true, i.e.,

$$\frac{D_{M_0} - D_M}{\psi} \sim \chi^2_{p-q}.$$

   Again, for models other than the normal (with known dispersion), the χ^2 distribution is an asymptotically-valid approximation to the exact distribution of the test statistic.

Expressions for deviance for a number of commonly-used models appear in Table 1; note that for the normal model, the deviance is a familiar quantity, the sum of the squared residuals.

Predicted Values, Residuals and Diagnostics. Predicted values and residuals are critical to diagnosing lack of model fit in ordinary regression models, and this is also the case for the broader class of GLMs. Residuals for the normal regression model are simply $y_i - \hat{\mu}_i$, but for the broader set of models with link functions other than the identity, predicted values and residuals are more ambiguous. It is helpful to make a distinction between prediction on the scale of the linear predictors $\eta_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$ and prediction on the scale of the observed response y_i, for which $E(y_i) = \mu_i = g^{-1}(\eta_i)$. In turn, it is possible to define *response residuals* $r_i = y_i - g^{-1}(\mathbf{x}_i\hat{\boldsymbol{\beta}})$. However, unlike the residuals from linear regression models, the response residuals for GLMs are not guaranteed to have the useful properties of ordinary regression residuals (e.g., summing to zero within a given data set, and having mean zero and normal distributions across repeated sampling).

A number of other types of residuals for GLMs have been proposed in the literature McCullagh and Nelder (1989, §2.4), including *deviance residuals*, defined as r_i^D = sign(y_i - $\hat{\mu}_i)\sqrt{d_i}$, where d_i is the observation-specific contributions to the deviance statistic D_M defined above. Deviance residuals have a number of useful properties: (a) they increase in magnitude with y_i - $\hat{\mu}_i$; (b) $\sum(r_i^D)^2 = D_M$; (c) have close to a normal distribution with mean zero and standard deviation one, irrespective of the type of GLM employed.

Outlier detection and other measures of influence are important data analytic tools, pioneered in the context of linear regression, but have been extended to GLMs; see Lindsey (1997, 227) and Hastie and Pregibon (1992, §6.3.4) for examples and references.

Extensions: GLMs have been extended and elaborated in numerous directions:

- **Non-constant variance**. The discussion so far has focused on GLMs where var(y_i) is constant (e.g., normal and binomial). But GLMs can be used for models such as the Poisson, where the $E(y_i)$ = var(y_i), extensions to such as the negative binomial McCullagh and Nelder (1989, §6.2.3), and variants such as the gamma GLM for non-negative continuous variables where var(y_i) increases with the square of the mean of y_i McCullagh and Nelder (1989, ch8).

- **Modeling dispersion**. In the usual setup, covariates $\mathbf{x}_i$ are used to model the conditional mean of y_i, $\mu_i = g^{-1}(\mathbf{x}_i\boldsymbol{\beta})$, with var($y_i$) = $\psi V(\mu_i)$. When substantive interest focuses on also modeling the variance of y_i, the GLM approach can be extended by letting the dispersion parameter ψ vary across observations conditional on covariates, typically via a gamma GLM. Any joint modeling of mean effects and dispersion effects is fraught with difficulty McCullagh and Nelder (1989, §10.3): the mean and dispersions models necessarily interact with one another, such that misspecification of one can be captured by the other, with the risk of misleading conclusions about where particular covariates have their impact on y_i (in the mean or variance).

- **Generalizing the link function**. Theory is usually silent as to the mapping from the linear predictors $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$ to $E(y_i) = \mu_i$. For many GLMs it is straightforward to specify a family of links indexed by one or more parameters to be estimated (McCullagh and Nelder, 1989, §11.3). For instance, for a binomial GLM, rather than assume either a logit or probit inverse-link, one could use the CDF of *t*-distribution with unknown degrees of freedom ν, with logit corresponding to $\nu \approx 8$ and probit when $\nu > 30$ (e.g., Albert and Chib, 1993).

Other references. While McCullagh and Nelder (1989) is the standard reference for GLMs, other treatments include Fahrmeir and Tutz (1994) and Lindsey (1997). Many specific treatments of the analysis of categorical variables make extensive use of the GLM approach, such as Collet (1991). Hastie and Pregibon (1992) and Venables and Ripley (2002, Ch7) discuss the implementation of GLMs in S-Plus and R.

References

Albert, James H. and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669--79.

Collet, D. 1991. *Modelling Binary Data*. London: Chapman and Hall.

Fahrmeir, F. and G. Tutz. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag.

Hastie, Trevor J. and Daryl Pregibon. 1992. "Generalized Linear Models." In *Statistical Models in S*, ed. John M. Chambers and Trevor J. Hastie. Pacific Grove, California: Wadsworth and Brooks/Cole Chapter 6.

Lindsey, James K. 1997. *Applying Generalized Linear Models*. New York: Springer.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. Second ed. London: Chapman and Hall.

Nelder, J.A. and R.W.M. Wedderburn. 1972. "Generalized linear models." *Journal of the Royal Statistical Society, Series A* 135:370--84.

Venables, William N. and Brian D. Ripley. 2002. *Modern Applied Statistics with S-PLUS*. Fourth ed. New York: Springer-Verlag.