GENERALIZED LINEAR MODELS

Joseph M. Hilbe
Arizona State University

1. HISTORY

Generalized Linear Models (GLM) is a covering algorithm allowing for the estimation of a number of otherwise distinct statistical regression models within a single framework. First developed by John Nelder and R.W.M. Wedderburn in 1972, the algorithm and overall GLM methodology has proved to be of substantial value to statisticians in terms of the scope of models under its domain as well as the number of accompanying model statistics facilitating an analysis of fit. In the early days of statistical computing - from 1972 to 1990 - the GLM estimation algorithm also provided a substantial savings of computing memory compared to what was required using standard maximum likelihood techniques. Prior to Nelder and Wedderburn's efforts, GLM models were typically estimated using a Newton-Raphson type full maximum likelihood method, with the exception of the Gaussian model. Commonly known as normal or linear regression, the Gaussian model is usually estimated using a least squares algorithm. GLM, as we shall observe, is a generalization of ordinary least squares regression, employing a weighted least squares algorithm that iteratively solves for parameter estimates and standard errors.

In 1974, Nelder coordinated a project to develop a specialized statistical application called GLIM, an acronym for Generalized Linear Interactive Modeling. Sponsored by the Royal Statistical Society and Rothamsted Experimental Station, GLIM provided the means for statisticians to easily estimate GLM models, as well as other more complicated models which could be constructed using the GLM framework. GLIM soon became one of the most used statistical applications worldwide, and was the first major statistical application to fully exploit the PC environment in 1981. However, it was discontinued in 1994. Presently, nearly all leading general purpose statistical packages offer GLM modeling capabilities; e.g. SAS, R, Stata, S-Plus, Genstat, and SPSS.

2. THEORY

Generalized linear models software, as we shall see, allows the user to estimate a variety of models from within a single framework, as well as providing the capability of changing models with minimal effort. GLM software also comes with a host of standard residual and fit statistics, which greatly assist researchers with assessing the comparative worth of models.

Key features of a generalized linear model include 1) having a response, or dependent variable, selected from the single parameter exponential family of probability distributions, 2) having a link function that linearizes the relationship between the fitted value and explanatory predictors, and 3) having the ability to be estimated using an Iteratively Re-weighted Least Squares (IRLS) algorithm.

The exponential family probability function upon which GLMs are based can be expressed as

$$f(y_i;\theta_i,\phi) = \exp\{(y_i\theta_i - b(\theta_i))/\alpha_i(\phi) - c(y_i;\phi)\} \tag{1}$$

where the distribution is a function of the unknown data, y, for given parameters θ and ϕ. For generalized linear models, the probability distribution is re-parameterized such that the distribution is a function of unknown parameters based on known data. In this form the distribution is termed a likelihood function, the goal of which is to determine the parameters making the data most likely. Statisticians log-transform the likelihood function in order to convert it to an additive rather than the multiplicative scale. Doing so greatly facilitates estimation based on the function. The log-likelihood function is central to all maximum likelihood estimation algorithms. It is also the basis of the deviance function, which was traditionally employed in GLM algorithms as both the basis of convergence and as a goodness-of-fit statistic. The log-likelihood is defined as

$$L(\theta_i;y_i,\phi) = \sum_{i=1}^{n}\{(y_i\theta_i - b(\theta_i))/\alpha_i(\phi) - c(y_i;\phi)\} \tag{2}$$

where θ is the link function, $b(\theta)$ the cumulant, $\alpha_i(\phi)$ the scale, and $c(y;\phi)$ the normalization term, guaranteeing that the distribution sums to one. The first derivative of the cumulant with respect to θ, $b'(\theta)$, is the mean of the function, μ; the second derivative, $b''(\theta)$, is the variance, $V(\mu)$. The deviance function is given as

$$2\sum_{i=1}^{n}\left\{L(y_i;y_i) - L(y_i,\mu_i)\right\} \tag{3}$$

Table 1 presents the standard probability distribution functions (PDF) belonging to the GLM family.

Table 1. GLM families : canonical

FAMILY	CHARACTERISTICS
Continuous distributions	
Gaussian	standard normal or linear regression
Gamma	positive-only continuous
Inverse Gaussian	positive-only continuous
Count	
Poisson	equidispersed count
Negative binomial (NB-C)	count, with the ancillary parameter a constant
Binary - Bernoulli	binomial distribution with $m = 1$.
Logistic	binary (1/0) response
Binomial	proportional (y/m) : y = number of 1's
Logistic (grouped)	m = cases having same covariate pattern

Each of the distributions in Table 1 are members of the exponential family. It should be noted, however, that the three continuous GLM distributions are usually parameterized with two rather than one parameter: Gaussian, gamma, and inverse Gaussian. Within the GLM framework though, the scale parameter is not estimated, although it is possible to point-estimate the scale value from the dispersion statistic, which is typically displayed in GLM model output. Binomial and count models have the scale value set at 1.0. As a consequence, $\alpha(\phi)$ and ϕ are many times excluded when presenting the GLM-based exponential log-likelihood.

Table 2 provides the formulae for the deviance and log-likelihoods of each GLM family. Also provided is the variance for each family function. The first line of each GLM distribution or family shows the deviance, with the next two providing the log-likelihood functions parameterized in terms of μ and $x'\beta$ respectively. The $x'\beta$ parameterization is used when models are estimated using a full maximum likelihood algorithm.

Table 2. GLM variance, deviance, and log-likelihood functions

FAMILY	VARIANCE, DEVIANCE, LOG-LIKELIHOOD ($\mu\vert x\beta$)
Gaussian	$\sum(y-\mu)^2$
1	$\sum\{(y\mu-\mu^2/2)/\sigma^2 - y^2/2\sigma^2 - 5\ln(2\pi\sigma^2)\}$
	$\sum\{[y(x\beta)-(x\beta)^2/2]/\sigma^2 - y^2/2\sigma^2 - 5\ln(2\pi\sigma^2)\}$
Bernoulli	$2\sum\{y\ln(y/\mu)+(1-y)\ln((1-y)/(1-\mu))\}$
$\mu(1-\mu)$	$\sum\{y\ln(\mu/(1-\mu))+\ln(1-\mu)\}$
	$\sum\left\{y(x\beta)-\ln(1+\exp(x\beta))\right\}$
Binomial	
$\mu(1-\mu/m)$	$2\sum\{y\ln(y/\mu)+(m-y)\ln((m-y)/(m-\mu))\}$
	$\sum\{y\ln(\mu/m)+(m-y)\ln(1-\mu/m)+\ln\Gamma(m+1)-\ln\Gamma(y+1)$ $+\ln\Gamma(m-y+1)\}$
	$\sum\{y\ln((\exp(x\beta))/(1+\exp(x\beta)))-(m-y)\ln(\exp(x\beta)+1)$ $+\ln\Gamma(m+1)-\ln\Gamma(y+1)+\ln\Gamma(m-y+1)\}$
Poisson	$2\sum\{y\ln(y/\mu)-(y-\mu)\}$
μ	$\sum\left\{y\ln(\mu)-\mu-\ln\Gamma(y+1)\right\}$
	$\sum\left\{y(x\beta)-\exp(x\beta)-\ln\Gamma(y+1)\right\}$
NB2	$2\sum\{y\ln(y/\mu)-(y+1/\alpha)\ln((1+\alpha y)/(1+\alpha\mu))\}$
$\mu+\alpha\mu^2$	$\sum\{y\ln((\alpha\mu)/(1+\alpha\mu))-(1/\alpha)\ln(1+\alpha\mu)+\ln\Gamma(y+1/\alpha)$ $-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}$
	$\sum\{y\ln((\alpha\exp(x\beta))/(1+\alpha\exp(x\beta)))-(\ln(1+\alpha\exp(x\beta)))/\alpha$ $+\ln\Gamma(y+1/\alpha)-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}$
NBC	$2\sum\{y\ln(y/\mu)-(y+1/\alpha)\ln((1+\alpha y)/(1+\alpha\mu))\}$
$\mu+\alpha\mu^2$	$\sum\{y\ln(\alpha\mu/(1+\alpha\mu))-(1/\alpha)\ln(1+\alpha\mu)+\ln\Gamma(y+1/\alpha)$ $-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}$
	$\sum\{y(x\beta)+(1/\alpha)\ln(1-\exp(x\beta))+\ln\Gamma(y+1/\alpha)-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}$
Gamma	$2\sum\{(y-\mu)/\mu-\ln(y/\mu)\}$

μ^2	$\sum\{((y/\mu)+\ln(\mu))/-\phi+\ln(y)(1-\phi)/\phi-\ln(\phi)/\phi-\ln\Gamma(1/\phi)\}$
	$\sum\{(y(x\beta)-\ln(x\beta))/-\phi+\ln(y)(1-\phi)/\phi-\ln(\phi)/\phi-\ln\Gamma(1/\phi)\}$
Inv Gauss	$\sum\{(y-\mu)^2/(y\mu^2)\}$
μ^3	$\sum\{(y/(2\mu^2)-1/\mu)/-\sigma^2+1/(-2y\sigma^2)-5\ln(2\pi y^3\sigma^2)\}$
	$\sum\{y/(2x\beta)-\sqrt{x\beta}/-\sigma^2+1/(-2y\sigma^2)-5\ln(2\pi y^3\sigma^2)\}$

Note that the link and cumulant functions for each of the above GLM log-likelihood functions can easily be abstracted from the equations, which are formatted in terms of the exponential family form as defined in Equation 1. For example, the link and cumulant of the Bernoulli distribution, upon which logistic regression is based, are respectively $\ln(\mu/(1-\mu))$ and $-\ln(1-\mu)$. With the link function defined in this manner, the linear predictor for the canonical Bernoulli model (logit) is expressed as:

$$\theta_i = x_i'\beta = \ln(\mu_i/(1-\mu_i)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n. \tag{4}$$

In GLM terminology, $x'\beta$ is also referred to as η, and the link as $g(\mu)$. For links directly derived from the GLM family PDF, the following terms are identical:

$$\theta = x_i'\beta = \eta = g(\mu). \tag{5}$$

The link function may be inverted such that μ is defined in terms of η. The resulting function is called the inverse link function, or $g^{-1}(\eta)$. For the above logit link, $\eta = \ln(\mu/(1-\mu))$. μ is therefore defined, for each observation in the logistic model, as

$$\mu_i = 1/(1+\exp(-\eta_i)) = (\exp(\eta_i))/(1+\exp(\eta_i)) \tag{6}$$

or

$$\mu_i = 1/(1+\exp(-x_i'\beta)) = (\exp(x_i'\beta))/(1+\exp(x_i'\beta)) \tag{7}$$

Another key feature of generalized linear models is the ability to use the GLM algorithm to estimate non-canonical models; i.e. models in which the link function is not directly derived from the underlying pdf, i.e, $x'\beta$ or η is not defined in terms of the value of θ given in the above listing of log-likelihood functions. Theoretically any link function can be associated with a GLM log-likelihood, although many might not be appropriate for the given data. A power link is sometimes used for non-binomial models where the power p in μ^p is allowed to vary. The statistician employs a value for the power that leads to a minimal value for the deviance. Powers typically range from -3 to 2, with μ^2 being the square link, μ^1 the log, μ^0 the identity, and μ^{-1} and μ^{-2} the inverse and inverse quadratic link functions respectively. Intermediate links are also used; e.g. $\mu^{.5}$, the square root link. The normal linear model has an identity link, with the linear predictor being identical to the fitted value.

The probit and log-linked negative binomial (NB-2) models are two commonly used non-canonical linked regression models. The probit link is often used with the binomial distribution for probit models. Although the probit link is not directly derived from the binomial PDF, the estimates of the GLM-based probit model are identical to those produced using full maximum likelihood methods. The canonical negative binomial (NB-C) is not the traditional negative binomial used to model overdispersed Poisson data. Rather, the use of the log link with the negative binomial (LNB) family duplicates estimates produced by full maximum likelihood NB-2 commands. However, like all non-canonical models, the standard errors of the LNB are slightly different from those of a full maximum likelihood NB-2, unless the traditional GLM algorithm in Table 5 is amended to produce an observed information matrix that is characteristic of full maximum likelihood estimation. The information derived from the algorithm given in Table 5 uses an expected information matrix, upon which standard errors are based. Applications such as Stata's *glm* command, SAS's *Genmod* procedure, and R's *glm*() and *glm.nb*() functions allow the user to select which information is to be used for standard errors.

The negative binomial family was not added to commercial GLM software until 1993 (Stata), and is in fact a member of the GLM family only if its ancillary or heterogeneity, parameter is entered into the algorithm as a constant. Setting the ancillary parameter, α, to a value that minimizes the Pearson dispersion statistic closely approximates the value of α estimated using a full maximum likelihood command. SAS, Stata, and R provide the capability for a user to estimate α using a maximum likelihood subroutine, placing the value determined into the GLM algorithm as a constant. The resulting estimates and standard errors are identical to a full NB-2 estimation. These applications also provide the capability of allowing the software to do this automatically.

The ability to incorporate non-canonical links into GLM models greatly extends the scope of models which may be estimated using its algorithm. Commonly used non-canonical models are shown in Table 3.

Table 3. Foremost Non-Canonical Models

FAMILY-LINK	FUNCTION
Continuous Distributions	
Lognormal	Positive continuous
Log-gamma	Exponential survival model
Log-inverse Gaussian	Steep initial peak; long slow tail
Bernoulli/Binomial:	
Probit	normal
Complementary loglog	asymmetric distribution: > 0.5 elongated
Loglog	asymmetric distribution: < 0.5 elongated
Negative Binomial	
Log (NB2)	overdispersed Poisson

The link, inverse link, and first derivative of the link for the canonical functions of the standard GLM families, as well as the most used non-canonical functions, are given in Table 4.

Table 4, GLM link functions (* canonical)

LINK NAME	LINK	INVERSE LINK	1^{st} DERIVATIVE
Gaussian			
*Identity	μ	η	1
Binomial (Bernoulli: $m = 1$)			
*Logit	$\ln(\mu/(m-\mu))$	$m/(1+\exp(-\eta))$	$m/(\mu(m-\mu))$
Probit	$\Phi^{-1}(\mu/m)$	$m\Phi(\eta)$	$m/\phi\{\Phi^{-1}(\mu/m)\}$
Cloglog	$\ln(-\ln(1-\mu/m))$	$m(1-\exp(-\exp(\eta)))$	$(m(1-\mu/m)\ln(1-\mu/m))^{-1}$
Poisson			
*Log	$\ln(\mu)$	$\exp(\eta)$	$1/\mu$
Neg Bin			
*NB-C	$\ln(\mu/(\mu+1/\alpha))$	$\exp(\eta)/(\alpha(1-\exp(\eta)))$	$1/(\mu+\alpha\mu^2)$
Log	$\ln(\mu)$	$\exp(\eta)$	$1/\mu$
Gamma			
*Inverse	$1/\mu$	$1/\eta$	$-1/\mu^2$
Inverse Gaussian			
*Inv Quad	$1/\mu^2$	$1/\sqrt{\eta}$	$-1/\mu^3$

3. IRLS ALGORITHM

Generalized linear models have traditionally been modeled using an Iteratively Re-Weighted Least Squares (IRLS) algorithm. IRLS is a version of maximum likelihood called Fisher Scoring, and can take a variety of forms. A standard IRLS schematic algorithm is given in Table 5.

Table 5. Generic GLM Estimating Algorithm (Expected Information Matrix)

```
$\mu = (y + \text{mean}(y))/2$               //  initialize $\mu$; non-binomial
$\mu = (y + 0.5)/(n + 1)$                    //  initialize $\mu$; binomial
$\eta = g(\mu)$                              //  initialize $\eta$; link
WHILE (abs($\Delta$ Dev)>tolerance){         //  loop
  $w = 1/(Vg'^2)$                            //  weight
  $z = \eta + (y - \mu)g'$                   //  working response
  $\beta = (X'wX)^{-1}X'wz$                  //  estimation of parameters
  $\eta = x'\beta$                           //  linear predictor,$\eta$
  $\mu = g^{-1}(\eta)$                       //  fit, $\mu$; inverse link
  Dev0 = Dev
  Dev = Deviance function                    //  deviance or LL
  $\Delta$ Dev = Dev-Dev0                    //  check for difference
}
```

$\text{Chi2} = \sum(y-\mu)^2/V(\mu)$ // Pearson χ^2
$\text{AIC} = (-2LL + 2p)/n$ // AIC GOF statistic
$\text{BIC} = -2 \cdot LL + p \cdot \ln n$ // BIC GOF statistic
Where $p =$ number of model predictors + intercept
$n =$ number of observations in model
$LL =$ log-likelihood function
$V =$ variance; $g(\mu) =$ link; $g^{-1}(\eta) =$ inverse link; $g' = \partial\eta/\partial\mu$

4. GOODNESS-OF-FIT

GLM models are traditionally evaluated as to their fit based on the deviance and Pearson Chi2, or χ^2, statistics. Lower values of these statistics indicate a better fitted model. Recently, statisticians have also employed the Akaike (AIC) and Bayesian (BIC) Information Criterion statistics as measures of fit. Lower values of the AIC and BIC statistics also indicate better fitted models. The Pearson Chi2, AIC, and BIC statistics are defined in Table 5, and are calculated after a model has been estimated.

The Pearson dispersion statistic is used with Poisson, negative binomial, and binomial models as an indicator of excessive correlation in the data. Likelihood based models, being derived from a PDF, assume that observations are independent. When they are not, correlation is observed in the data. Values of the Pearson dispersion greater than 1.0 indicate more correlation in the data than is warranted by the assumptions of the underlying distribution. Some statisticians have used the deviance statistic on which to base the dispersion, but simulation studies have demonstrated that Pearson is the correct statistic. See **Modeling count data** in this Encyclopedia for additional information.

From the outset, generalized linear models software has offered users a number of useful residuals which can be used to assess the internal structure of the modeled data. Pearson and deviance residuals are the two most recognized GLM residuals associated with GLM software. Both are observation-based statistics, providing the proportionate contribution of an observation to the overall Pearson Chi2 and deviance fit statistics. The two residuals are given, for each observation, as:

$$\text{Pearson} \qquad (y-\mu)/\sqrt{V(\mu)} \tag{8}$$

$$\text{deviance} \qquad \text{sgn}(y-\mu)\sqrt{\text{deviance}} \tag{9}$$

The Pearson Chi2 and deviance fit can also be calculated on the basis of their residuals by taking the square of each of the residuals respectively, and summing them over all observations in the model. However, they are seldom calculated in such a manner.

Both the Pearson and deviance residuals are usually employed in standardized form. The standardized versions of the Pearson and deviance residuals are given by dividing the respective statistic by $\sqrt{1-h}$ where h is the hat matrix diagonal. Standardized Pearson residuals are normalized to a standard deviation of 1.0 and are adjusted to account for the correlation between y and μ. The standardized deviance residuals are the most commonly used residuals for assessing the internal shape of the modeled data.

Another residual now finding widespread use is the Anscombe residual. First implemented into GLM software in 1993, it now enjoys use in many major software applications. The Anscombe residuals are defined specifically for each family, with the intent of normalizing the residuals as much as possible. The general formula for Anscombe residuals is given as

$$\int_y^\mu \mathrm{d}\mu V^{-1/3}(\mu) \tag{10}$$

with $V^{-1/3}(\mu)$ as the inverse cube of the variance. The Anscombe residual for the binomial family is displayed as

$$\frac{A(y) - A(\mu)}{\mu(1-\mu)^{-1/6}\sqrt{\dfrac{1-h}{m}}} \tag{11}$$

with $A()$ equal to $2.05339 \cdot$ (Incomplete Beta $(2/3, 2/3, z)$, z taking the value of μ or y. A standard use of this statistic is to graph it on either the fitted value, or the linear predictor. Values of the Anscombe residual are close to those of the standardized deviance residuals.

5. APPLICATION

Consider data from the 1912 Titanic disaster. Information was collected on the survival status, gender, age, and ticket class of the various passengers. With *age* (1=adult; 0=child) and *sex* (1=male; 0-female), and *class* (1=1$^{\text{st}}$; 2=2$^{\text{nd}}$; 3=3$^{\text{rd}}$) with 3$^{\text{rd}}$ class as the reference, a simple binary logistic regression can be run using a GLM command (Stata). The type of model to be estimated is declared using the $family()$ and $link()$ functions. $eform$ indicates that the coefficients are to be exponentiated, resulting in odds ratios for the logistic model. Note the fact that 1st class passengers had a near 6 times greater odds of survival than did 3rd class passengers. The statistics displayed in the model output are fairly typical of that displayed in GLM software applications.

```
.glm survived age sex class1 class2, family(bin) link(logit) eform
```

```
Generalized linear models No. of obs = 1316
Optimization : ML Residual df = 1313
. Scale parameter = 1
Deviance = 1276.200769 (1/df) Deviance = .973456
Pearson = 1356.674662 (1/df) Pearson = 1.03484
Variance function: V(u) = u . (1 - u) (Bernoulli)
Link function: g(u) = ln(u/(1 - u) (Logit)

. AIC = .9773562
Log likekihood = -638.1003845 BIC = .8139.863

=================================================================================
survived | Odds ratio OIM Std. Err. z P>|z| [95% Conf. Interval]
=================================================================================
. age | .3479809 .0844397 -4.35 0.000 .2162749 .5598924
. sex | .0935308 .0135855 -16.31 0.000 .0703585 .1243347
. class1 | 5.84959 .9986265 10.35 0.000 4.186109 8.174107
. class2 | 2.129343 .3731801 4.31 0.000 1.510315 3.002091
===================================================================================
```

Using R, the same model would be specified by
glm(survived $\sim$ age + sex + class1 + class2, family=binomial, link=logit, data=titanic).

References

[1] Collett, D. (2003). *Modeling Binary Data*, 2nd edition, London: Chapman & Hall/CRC.

[2] Dobson, A. J. and A. G. Barnett (2008). *An Introduction to Generalized Linear Models*, 3rd edition, Boca Raton, FL: Chapman & Hall/CRC.

[3] Faraway, J.J. (2006). *Extending the Linear Model with R*, Boca Raton: FL: Chapman & Hall/CRC.

[4] Hardin, J. W. and J.M. Hilbe (2007) *Generalized Linear Models and Extensions*, 2nd edition, College Station, TX: Stata Press.

[5] Hilbe, J. M. (1994). Generalized Linear Models, *The American Statistician* **48**: 255–265.

[6] Hilbe, J. M. (2007). *Negative Binomial Regression*, Cambridge: Cambridge University Press.

[7] Hilbe, J. M. (2009). *Logistic Regression Models*, Boca Raton, FL: Chapman & Hall/CRC.

[8] Hoffmann, J. P. (2004). *Generalized Linear Models; an applied approach*, Boston: Allyn and Bacon.

[9] McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, 2nd edition, London; Chapman & Hall/CRC.

[10] Nelder, J. A. and R.W.M. Wedderburn (1972). Generalized Linear Models, *Journal of the Royal Statistical Society*, Series A **135**: 370–384.

[11] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**: 439–447.

Reprinted with permission from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC