Generalized Linear Models

Jeff Gill

Department of Political Science, University of Florida

234 Anderson Hall, PO Box 117325, Gainesville, FL 32611-7325

Voice: 352-392-0262x272, Fax: 352-392-8127, Email: jgill@polisci.ufl.edu

Word Count: 2846.

Introduction to Generalized Linear Models

Generalized linear models expand the basic structure of the well-known linear model to accommodate non-normal and non-interval measured outcome variables in a single unified theoretical form. It is common in the social sciences to encounter outcome variables that do not fit the standard assumptions of the linear model, and as a result many distinct forms of regression have been developed: Poisson regression for counts as outcomes, logit and probit for dichotomous explanations, exponential models for durations, gamma models for truncated data, and more. While these forms are commonly used and well-understood, it is not widely known among practitioners that nearly all of these particularistic regression techniques are special forms of the *generalized linear model*.

With the generalized linear model, explanatory variables are treated in exactly the usual fashion by creating a linear systematic component, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ which defines the right-hand-side of the statistical model. Since the expectation of the outcome variable, $\boldsymbol{\mu} = \bar{\mathbf{y}}$ is no longer from an interval-measured form, standard central limit and asymptotic theory no longer applies in the same way and a "link" function is required to define the relationship between the linear systematic component of the

data and the mean of the outcome variable, $\boldsymbol{\mu} = g(\mathbf{X}\boldsymbol{\beta})$. If the specified link function is simply the identity function, $g(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$, then the generalized linear model reduces to the linear model, hence its name.

The generalized linear model is based on well-developed theory, starting with Nelder and Wedderburn (1972) and McCullagh and Nelder (1989), which states that any parametric form for the outcome variable that can be recharacterized (algebraically) into the *exponential family form* leads to a link function that connects the mean function of this parametric form to the linear systematic component. This exponential family form is simply a standardized means of expressing various probability mass functions (PMFs, for discrete forms) and probability density functions (PDFs, for continuous forms).

The value of the GLM approach is that a seemingly disparate set of nonlinear regression models can be integrated into a single framework where the processes of: specification searching, numerical estimation, residuals analysis, goodness of fit analysis, and reporting, are all unified. Thus researchers and practitioners can learn a singularly general procedure that fits a wide class of data types and can also be developed to include even broader classes of assumptions than those described here.

The Exponential Family Form

The key to understanding the generalized linear model is knowing how common probability density functions for continuous data forms and probability mass functions for discrete data forms can be expressed in exponential family form; a form that readily produces link functions and moment statistics. The exponential family form originates with Fisher (1934, p.288-96), but it was not shown until later that

members of the exponential family have fixed dimension sufficient statistics, and produce unique maximum likelihood estimates with strong consistency and asymptotic normality (Gill 2000, p.22).

Consider the one-parameter conditional PDF or PMF for the random variable Z of the form: $f(z|\zeta)$. It is classified as being an exponential family form if it can be expressed as:

$$f(z|\zeta) = \exp\big[\ \underbrace{t(z)u(\zeta)}_{\substack{\text{interaction}\\ \text{component}}} + \underbrace{\log(r(z)) + \log(s(\zeta))}_{\text{additive component}}\ \big]. \tag{1}$$

given that r and t are real-valued functions of z that do not depend on ζ, and s and u are real-valued functions of ζ that do not depend on z, with $r(z) > 0, s(\zeta) > 0\ \forall z, \zeta$. The key feature is the distinctiveness of the subfunctions within the exponent. The piece labeled *interaction component* must have $t(z)$, strictly a function of z, multiplying $u(\zeta)$, strictly a function of ζ. Furthermore, the *additive component* must have similarly distinct, but now summed, subfunctions with regard to z and ζ.

The canonical form of the exponential family expression is a simplification of (1) that reveals greater structure and gives a more concise summary of the data. This one-to-one transformation works as follows. Make the transformations $y = t(z)$ and $\theta = u(\zeta)$ if necessary (i.e. they are not already in canonical form), and now express (1) as:

$$f(y|\theta) = \exp\big[y\theta - b(\theta) + c(y)\big]. \tag{2}$$

The subfunctions in (2) have specific identifiers: y is the canonical form for the data (and typically a sufficient statistic as well), θ is the natural canonical form for the unknown parameter, and $b(\theta)$ is often called the "cumulant function" or "normalizing constant". The function $c(y)$ is usually not important in the estimation process. However, $b(\theta)$ plays a key role in both determination of the mean and variance

functions as well as the link function. It should also be noted that the canonical form is invariant to random sampling, meaning that it retains its functionality for iid (independent, identically distributed) data:

$$f(\mathbf{y}|\theta) = \exp\left[\sum_{i=1}^{n} y_i\theta - nb(\theta) + \sum_{i=1}^{n} c(y_i)\right],$$

and under the extension to multiple parameters through a k-length parameters vector $\boldsymbol{\theta}$:

$$f(y|\boldsymbol{\theta}) = \exp\left[\sum_{j=1}^{k} (y\theta_j - b(\theta_j)) + c(y)\right].$$

As an example of this theory, we can rewrite the familiar binomial PMF:

$$f(y|n,p) = \binom{n}{y} p^y (1-p)^{n-y}$$

in exponential family form:

$$f(y|n,p) = \exp\left[\underbrace{y\log\left(\frac{p}{1-p}\right)}_{y\theta} - \underbrace{(-n\log(1-p))}_{b(\theta)} + \underbrace{\log\binom{n}{y}}_{c(y)}\right].$$

Here the subfunctions are label to correspond to the canonical form previously identified.

The Exponential Family Form and Maximum Likelihood

In order to obtain estimates of the unknown k-dimensional $\boldsymbol{\theta}$ coefficient vector, given an observed matrix of data values: $f(\boldsymbol{\theta}|\boldsymbol{X})$, we can employ the standard technique of maximizing the likelihood function with regard to coefficient values to find the "most likely" values of the $\boldsymbol{\theta}$ vector (Fisher 1925, p.707-9). Asymptotic theory assures us that for sufficiently large samples the likelihood surface is unimodal in k

dimensions for exponential family forms, so the MLE process is equivalent to finding the k-dimensional mode.

Regarding $f(\mathbf{X}|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ for given observed (now fixed) data $\mathbf{X}$, then $L(\boldsymbol{\theta}|\mathbf{X}) = f(\mathbf{X}|\boldsymbol{\theta})$ is called a likelihood function. The maximum likelihood estimate (MLE), $\hat{\boldsymbol{\theta}}$, has the characteristic that $L(\hat{\boldsymbol{\theta}}|\mathbf{X}) \geq L(\boldsymbol{\theta}|\mathbf{X}) \ \forall \ \boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the admissible range of $\boldsymbol{\theta}$ given by the parametric form assumptions.

It is more convenient to work with the natural log of the likelihood function, and this does not change any of the resulting parameter estimates because the likelihood function and the log likelihood function have identical modal points. Using (2) with a scale parameter added, ψ, the likelihood function in canonical exponential family notation is:

$$\ell(\theta, \psi|\boldsymbol{y}) = \log(f(\boldsymbol{y}|\theta, \psi)) = \log\left(\exp\left[\frac{\boldsymbol{y}\theta - b(\theta)}{a(\psi)} + c(\boldsymbol{y}, \psi)\right]\right) = \frac{\boldsymbol{y}\theta - b(\theta)}{a(\psi)} + c(\boldsymbol{y}, \psi). \tag{3}$$

Since all of the terms are expressed in the exponent of the exponential family form, removing this exponent through the log of the likelihood gives a very compact form. The score function is first derivative of the log likelihood function with respect to the parameters of interest. For the time being the scale parameter, ψ, is treated as a *nuisance parameter* (not of primary interest), and we estimate a scalar θ. The resulting score function, denoted as $\dot{\ell}(\theta|\psi, \boldsymbol{y})$, is:

$$\dot{\ell}(\theta|\psi, \boldsymbol{y}) = \frac{\partial}{\partial\theta}\ell(\theta|\psi, \boldsymbol{y}) = \frac{\partial}{\partial\theta}\left[\frac{\boldsymbol{y}\theta - b(\theta)}{a(\psi)} + c(\boldsymbol{y}, \psi)\right] = \frac{\boldsymbol{y} - \frac{\partial}{\partial\theta}b(\theta)}{a(\psi)} \tag{4}$$

The mechanics of the maximum likelihood process involve setting $\dot{\ell}(\theta|\psi, \boldsymbol{y})$ equal to zero then solving for the parameter of interest giving the MLE: $\hat{\theta}$. Furthermore, he *Likelihood Principle* states that once the data are observed, all of the available evidence for estimating $\hat{\theta}$ is contained in the calculated likelihood function, $\ell(\theta, \psi|\boldsymbol{y})$.

The value of identifying the $b(\theta)$ function lies in its direct link to the mean and variance functions. The expected value calculation of (2) with respect to the data (Y) is in the notation of (3) is:

$$E_Y\left[\frac{y-\frac{\partial}{\partial\theta}b(\theta)}{a(\psi)}\right]=0$$

$$\int_Y yf(y)dy-\int_Y\frac{\partial b(\theta)}{\partial\theta}f(y)dy=0$$

$$\underbrace{\int_Y yf(y)dy}_{E[Y]}-\frac{\partial b(\theta)}{\partial\theta}\underbrace{\int_Y f(y)dy}_{1}=0 \qquad \Longrightarrow \quad E[Y]=\frac{\partial}{\partial\theta}b(\theta),$$

where the second to last step requires general regularity conditions with regard to the bounds of integration and all exponential family distributions meet this requirement (Gill 2000, p.23). This means that (2) gives the mean of the associated exponential family of distributions: $\mu = b'(\theta)$. It can also be shown (Gill 2000, p.26) that the second derivative of $b(\theta)$ is the variance function. Then multiplying this by the scale parameter (if appropriate), and substituting back in for the canonical parameter produces the standard variance calculation. These calculations are summarized for the binomial PMF:

Table 1: Binomial Mean and Variance Functions

$b(\theta)$	$n\log\left(1+\exp(\theta)\right)$
$E[Y]=\frac{\partial}{\partial\theta}b(\theta)\Big\vert_{\theta=u(\zeta)}$	np
$\frac{\partial^2}{\partial\theta^2}b(\theta)$	$n\exp(\theta)(1+\exp(\theta))^{-2}$
$Var[Y]=a(\psi)\frac{\partial^2}{\partial\theta^2}b(\theta)\Big\vert_{\theta=u(\zeta)}$	$np(1-p)$

The Generalized Linear Model Theory

Start with the standard linear model meeting the Gauss-Markov conditions:

$$\underset{(n\times 1)}{\boldsymbol{V}} = \underset{(n\times p)(p\times 1)}{\boldsymbol{X\beta}} + \underset{(n\times 1)}{\boldsymbol{\epsilon}}$$

$$\underset{(n\times 1)}{E(\boldsymbol{V})} = \underset{(n\times 1)}{\boldsymbol{\theta}} = \underset{(n\times p)(p\times 1)}{\boldsymbol{X\beta}} \tag{5}$$

The right-hand sides of the two equations in (5) contain: $\boldsymbol{X}$, the matrix of observed data values, $\boldsymbol{X\beta}$, the "linear structure vector", and $\boldsymbol{\epsilon}$, the error terms. The left-hand side contains: $E(\boldsymbol{V}) = \boldsymbol{\theta}$, the vector of means: i.e. the systematic component. The variable, $\boldsymbol{V}$, is distributed iid normal with mean $\boldsymbol{\theta}$, and constant variance σ^2. Now suppose we generalize this with a new "linear predictor" based on the mean of the outcome variable, which is no longer required to be normally distributed or even continuous:

$$\underset{(n\times 1)}{g(\boldsymbol{\mu})} = \underset{(n\times 1)}{\boldsymbol{\eta}} = \underset{(n\times p)(p\times 1)}{\mathbf{X\beta}} \quad .$$

It is important here that $g()$ be an invertible, *smooth* function of the mean vector $\boldsymbol{\mu}$.

The effect of the explanatory variables is now expressed in the model only through the link from the linear structure, $\mathbf{X\beta}$, to the linear predictor, $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, controlled by the form of the link function, $g()$. This link function connects the linear predictor to the *mean* of the outcome variable not directly to the expression of the outcome variable itself, so the outcome variable can now take on a variety of non-normal forms. The link function connects the stochastic component which describes some response variable from a wide variety of forms to all of the standard normal theory

supporting the linear systematic component through the mean function:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$
$$g^{-1}(g(\boldsymbol{\mu})) = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\mu} = E(\mathbf{Y}).$$

Furthermore, this linkage is provided by the form of the mean function from the exponential family subfunction: $b(\theta)$.

For example, with the binomial PMF we saw that $b(\theta) = n\log(1+\exp(\theta))$, and $\theta = \log\left(\frac{p}{1-p}\right)$. Re-expressing this in link function notation is just $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \log\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right)$, or we could give the inverse link, $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\theta})}{1+\exp(\boldsymbol{\theta})}$.

The generalization of the linear model as described now has the following components:

I. **Stochastic Component:** $\boldsymbol{Y}$ is the random or stochastic component which remains distributed iid according to a specific exponential family distribution with mean $\boldsymbol{\mu}$.

II. **Systematic Component:** $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ is the systematic component with an associated Gauss-Markov normal basis.

III. **Link Function:** the stochastic component and the systematic component are linked by a function of $\boldsymbol{\eta}$ which is taken from the inverse the of the canonical link, $b(\theta)$.

IV. **Residuals:** Although the residuals can be expressed in the same manner as in the standard linear model, observed outcome variable value minus predicted outcome variable value, a more useful quantity is the deviance residual described below.

Estimating Generalized Linear Models

Unlike the standard linear model, estimating generalized linear models is not done with a closed-form analytical expression (i.e. $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$). Instead the maximum likelihood estimate of the unknown parameter is found with a weighted numerical process that repeatedly increases the likelihood with improved weights on each cycle: *iterative weighted least squares* (IWLS). This process, first proposed by Nelder and Wedderburn (1972, p.372-4) and first implemented in the `GLIM` software package, works for any GLM based on an exponential family form (and for some others). Currently, all professional-level statistic computing implementations now employ IWLS to numerically find maximum likelihood estimates for generalized linear models.

The overall strategy is to apply Newton-Raphson with Fisher Scoring to the normal equations, which is equivalent to iteratively applied, weighted *least squares* (and hence easy). Define the current (or starting) point of the linear predictor by:

$$\underset{(n\times 1)}{\hat{\boldsymbol{\eta}}_0} = \underset{(n\times p)(p\times 1)}{\mathbf{X}'\boldsymbol{\beta}_0}$$

with fitted value $\hat{\boldsymbol{\mu}}_0$ from $g^{-1}(\hat{\boldsymbol{\eta}}_0)$. Form the "adjusted dependent variable" according to:

$$\underset{(n\times 1)}{\mathbf{z}_0} = \underset{(n\times 1)}{\hat{\boldsymbol{\eta}}_0} + \underset{diag(n\times n)}{\left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}\Big|_{\hat{\boldsymbol{\mu}}_0}\right)} \underset{(n\times 1)}{(\mathbf{y} - \hat{\boldsymbol{\mu}}_0)}$$

which is a linearized form of the link function applied to the data. As an example of this derivative function, the binomial form looks like:

$$\boldsymbol{\eta} = \log\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right) \implies \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} = (\boldsymbol{\mu})^{-1}(1-\boldsymbol{\mu})^{-1}.$$

Now form the *quadratic weight matrix*, which is the variance of $\mathbf{z}$:

$$\underset{diag(n\times n)}{\boldsymbol{\omega}_0^{-1}} = \underset{diag(n\times n)}{\left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}\Big|_{\hat{\boldsymbol{\mu}}_0}\right)^2} \underset{diag(n\times n)}{v(\boldsymbol{\mu})|_{\hat{\boldsymbol{\mu}}_0}}$$

where $v(\boldsymbol{\mu})$ is the variance function: $\frac{\partial}{\partial \boldsymbol{\theta}} b'(\boldsymbol{\theta}) = b''(\boldsymbol{\theta})$. Also note that this process is necessarily iterative because both $\mathbf{z}$ and $\boldsymbol{\omega}$ depend on the current fitted value, $\boldsymbol{\mu}_0$. The general scheme can now be summarized in the three steps:

I. Construct $\mathbf{z}$, $\boldsymbol{\omega}$. Regress $\mathbf{z}$ on the covariates with weights to get a new interim estimate:

$$\underset{(p\times 1)}{\hat{\boldsymbol{\beta}}_1} = (\underset{(p\times n)}{\mathbf{X}'} \underset{(n\times n)}{\boldsymbol{\omega}_0} \underset{(n\times p)}{\mathbf{X}})^{-1} \underset{(p\times n)}{\mathbf{X}'} \underset{(n\times n)}{\boldsymbol{\omega}_0} \underset{(n\times 1)}{\mathbf{z}_0}$$

II. Use the coefficient vector estimate to update the linear predictor:

$$\hat{\boldsymbol{\eta}}_1 = \mathbf{X}'\hat{\boldsymbol{\beta}}_1$$

III. Iterate:

$$\begin{aligned} \mathbf{z}_1, \boldsymbol{\omega}_1 &\implies \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\eta}}_2 \\ \mathbf{z}_2, \boldsymbol{\omega}_2 &\implies \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\eta}}_3 \\ \mathbf{z}_3, \boldsymbol{\omega}_3 &\implies \hat{\boldsymbol{\beta}}_4, \hat{\boldsymbol{\eta}}_4 \\ &\vdots \end{aligned}$$

Under very general conditions, satisfied by the exponential family of distributions, the iterative weighted least squares procedure finds the mode of the likelihood function, thus producing the maximum likelihood estimate of the unknown coefficient vector, $\hat{\beta}$. Furthermore, the matrix produced by: $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}$ converges in probability to the variance matrix of $\hat{\beta}$ as desired.

Residuals and Deviances

One significant advantage of the generalized linear model the freedom it provides from the standard Gauss-Markov assumption that the residuals have mean zero and constant variance. Unfortunately this freedom comes with the price of interpreting more complex stochastic structures. The *response* residual vector, $\boldsymbol{R}_{Response} = \boldsymbol{Y} - \boldsymbol{X\beta}$, calculated for linear models can be updated to include the GLM link function, $\boldsymbol{R}_{Response} = \boldsymbol{Y} - g^{-1}(\boldsymbol{X\beta})$, but this does not then provide the nice distribution theory we get from the standard linear model.

A more useful, but related, idea for generalized linear models is the *deviance function*. This is built in a similar fashion to the likelihood ratio statistic, comparing the log likelihood from a proposed model specification to the maximum log likelihood possible through the saturated model (n data points, n specified parameters, using the exact same data and link function). The resulting difference is multiplied by two and called the summed deviance (Nelder and Wedderburn 1972, p.374-6). The goodness of fit intuition is derived from the idea that this sum constitutes the summed contrast of individual likelihood contributions with the native data contributions to the saturated model. The point here is to compare the log likelihood for the proposed model:

$$\ell(\hat{\boldsymbol{\theta}}, \psi|\boldsymbol{y}) = \sum_{i=1}^{n} \frac{y_i\hat{\boldsymbol{\theta}} - b(\hat{\boldsymbol{\theta}})}{a(\psi)} + c(\boldsymbol{y}, \psi)$$

to the same log likelihood function with identical data and the same link function, except that it now with n coefficients for the n data points, i.e. the saturated model log likelihood function:

$$\ell(\tilde{\boldsymbol{\theta}}, \psi|\boldsymbol{y}) = \sum_{i=1}^{n} \frac{y_i\tilde{\boldsymbol{\theta}} - b(\tilde{\boldsymbol{\theta}})}{a(\psi)} + c(\boldsymbol{y}, \psi).$$

The latter is the highest possible value for the log likelihood function achievable with the given data. The deviance function is then given by:

$$D(\boldsymbol{\theta}, \boldsymbol{y}) = 2\sum_{i=1}^{n}\left[\ell(\tilde{\boldsymbol{\theta}}, \psi|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}, \psi|\boldsymbol{y})\right] = 2\sum_{i=1}^{n}\left[y_i(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) - (b(\tilde{\boldsymbol{\theta}}) - b(\hat{\boldsymbol{\theta}})\right] a(\psi)^{-1}.$$

This statistic is asymptotically χ^2 with $n-k$ degrees of freedom (although high levels of discrete granularity in the outcome variable can make this a poor distributional assumption for datasets that are not so large). Fortunately, these deviance functions are commonly tabulated for many exponential family forms, and therefore do not require analytical calculation. In the binomial case the deviance function is:

$$D(m, p) = 2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \mu_i}\right)\right],$$

where the saturated log likelihood achieves the highest possible value for fitting $p_i = y_i/n_i$.

We can also look at the individual deviance contributions in an analogous way to linear model residuals. The single point deviance function is just the deviance function for the y_i^{th} point:

$$d(\boldsymbol{\theta}, y_i) = -2\left[y_i(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) - \left(b(\tilde{\boldsymbol{\theta}}) - b(\hat{\boldsymbol{\theta}}\right)\right] a(\psi)^{-1}.$$

A deviance residual at the y_i point is built upon this by:

$$R_{Deviance} = \frac{(y_i - \mu_i)}{|y_i - \mu_i|}\sqrt{|d(\boldsymbol{\theta}, y_i)|}$$

where $\frac{(y_i-\mu_i)}{|y_i-\mu_i|}$ is just a sign-preserving function.

Example: Multinomial Response Models

Consider a slight generalization of the running binomial example where instead of two possible events defining the outcome variable, there are now k events. The

outcome for an individual i is given as a $k-1$ length vector of all zeros except for a single one identifying the positive response: $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{i(k-1)}]$, or all zeros for an individual selecting the left-out reference category. It should be clear that this reduces to a binomial outcome for $k = 2$.

The objective is now to estimate the $k-1$ length of categorical probabilities for a sample size of n, $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\pi} = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_{k-1}]$, from the dataset consisting of the $n \times (k-1)$ outcome matrix $\mathbf{y}$ and the $n \times p$ matrix of p covariates $\mathbf{X}$ including a leading column of 1's. The PMF for this setup is now multinomial where the estimates are provided with a logit (but sometimes a probit) link function, giving for each of $k-1$ categories, the probability that the i^{th} individual picks category r:

$$P(\mathbf{y}_i = r|\mathbf{X}) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{k-1}\exp(\mathbf{X}\boldsymbol{\beta}_s)}$$

where $\boldsymbol{\beta}_r$ is the coefficient vector for the r^{th} category.

The data used are from the 1977 General Social Survey, and are a specific 3-category multinomial example analyzed by a number of authors. These are summarized in Table 2. Since there are only three categories in this application, the multinomial PMF for the i^{th} individual is given by:

$$f(\mathbf{y}_i|n_i, \boldsymbol{\pi}_i) = \frac{n_i}{y_{i1}!y_{i2}!(n - y_{i1} - y_{i2})!}\boldsymbol{\pi}_{i1}^{y_{i1}}\boldsymbol{\pi}_{i2}^{y_{i2}}(1 - \boldsymbol{\pi}_{i1} - \boldsymbol{\pi}_{i2})^{n-y_{i1}-y_{i2}}$$

and the likelihood is formed by the product across the sample. Is this an exponential family form such that we can treat this as a GLM problem? This PMF can be re-expressed as:

Table 2: Happiness and Schooling by Sex, 1977 GSS

		Years of Schooling			
	Self-Reported Status	< 12	12	13–16	17+
Male	Not happy	40	21	14	3
	Pretty happy	131	116	112	27
	Very happy	82	61	55	27
Female	Not happy	62	26	12	3
	Pretty happy	155	156	95	15
	Very happy	87	127	76	15

$$f(\mathbf{y}_i|n_i,\boldsymbol{\pi}_i) = \exp\Bigg[\Bigg(\underbrace{\left(\frac{\mathbf{y}_{i1}}{n_i},\frac{\mathbf{y}_{i2}}{n_i}\right)}_{\mathbf{y}_i'}\underbrace{\left(\log\left(\frac{\boldsymbol{\pi}_{i1}}{1-\boldsymbol{\pi}_{i1}-\boldsymbol{\pi}_{i2}}\right),\log\left(\frac{\boldsymbol{\pi}_{i2}}{1-\boldsymbol{\pi}_{i1}-\boldsymbol{\pi}_{i2}}\right)\right)'}_{\boldsymbol{\theta}_i}$$
$$-\underbrace{(-\log(1-\boldsymbol{\pi}_{i1}-\boldsymbol{\pi}_{i2}))}_{b(\theta_i)}\Bigg)n_i+\underbrace{\log\left(\frac{n_i}{y_{i1}!y_{i2}!(n-y_{i1}-y_{i2})!}\right)}_{c(y)}\Bigg]$$

where n_i is a weighting for the i^{th} case. This is clearly an exponential form, albeit now with a two-dimensional structure for $\mathbf{y}_i'$ and $\boldsymbol{\theta}_i$. The two-dimensional link function that results from this form is simply:

$$\boldsymbol{\theta}_i = g(\boldsymbol{\pi}_{i1},\boldsymbol{\pi}_{i2}) = \left(\log\left(\frac{\boldsymbol{\pi}_{i1}}{1-\boldsymbol{\pi}_{i1}-\boldsymbol{\pi}_{i2}}\right),\log\left(\frac{\boldsymbol{\pi}_{i2}}{1-\boldsymbol{\pi}_{i1}-\boldsymbol{\pi}_{i2}}\right)\right).$$

We can therefore interpret the results in the following way for a single respondent:

$$\log\left[\frac{P(\text{event 1})}{P(\text{reference category})}\right] = \log\left[\frac{\boldsymbol{\pi}_{i1}}{1-\boldsymbol{\pi}_{i1}-\boldsymbol{\pi}_{i2}}\right] = \mathbf{X}_i\boldsymbol{\beta}_1$$

$$\log\left[\frac{P(\text{event 2})}{P(\text{reference category})}\right] = \log\left[\frac{\boldsymbol{\pi}_{i2}}{1-\boldsymbol{\pi}_{i1}-\boldsymbol{\pi}_{i2}}\right] = \mathbf{X}_i\boldsymbol{\beta}_2$$

The estimated results for this model are contained in Table 3. What we observe from these results is that there is no evidence of gender effect (counter to other studies), and there is strong evidence of increased happiness for the three categories relative to the reference category of <12 years of school. Interesting, the Very Happy to Not Happy distinction increases with education, but the Pretty Happy to Not Happy distinction does not. The deviance residual is 8.68, indicating a good fit for 6 degrees of freedom (not in the tail of a χ^2 distribution), and therefore an improvement over the saturated model.

Table 3: Three-Category Multinomial Model Results

	Intercept	Female	12	13–16	17+
Pretty Happy	1.129	-0.181	0.736	1.036	0.882
	(0.148)	(0.168)	(0.196)	(0.238)	(0.453)
Very Happy	0.474	0.055	0.878	1.114	1.451
	(0.161)	(0.177)	(0.206)	(0.249)	(0.455)

References

Fisher, R. A. (1925). Theory of Statistical Estimation. *Proceedings of the Cambridge Philosophical Society* 22, 700-25.

Fisher, R. A. (1934). "Two New Properties of Mathematical Likelihood." *Proceedings of the Royal Society of London, A* 144, 285-307.

Gill, J. (2000). *Generalized Linear Models: A Unified Approach.* Thousand Oaks, CA: Sage.

McCullagh, P., and J. A. Nelder. (1989). *Generalized Linear Models.* Second Edition. New York: Chapman & Hall.

Nelder, J. A., and R. W. M. Wedderburn. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A* 135, 370-85.