Directed Reading in Time Series and Dynamic Analysis

April 2008

A Quick Introduction to Generalized Linear Models

The Exponential Family

We'll start with a random variable Z, which has realizations called z. We are typically interested in the conditional density (PDF) of Z, where the conditioning is on some parameter or parameters ψ:

$$f(z|\psi) = \Pr(Z = z|\psi)$$

A PDF is a member of the exponential family of probability functions if it can be written in the form:[1]

$$f(z|\psi) = r(z)s(\psi)\exp[q(z)h(\psi)] \tag{1}$$

where $r(\cdot)$ and $q(\cdot)$ are functions of z that do not depend on ψ, and (conversely) $s(\cdot)$ and $h(\cdot)$ are functions of ψ that do not depend on z. For reasons that will be apparent in a minute, it is also necessary that $r(z) > 0$ and $s(\psi) > 0$.

With a little bit of algebra, we can rewrite (1) as:

$$f(z|\psi) = \exp[\ln r(z) + \ln s(\psi) + q(z)h(\psi)]. \tag{2}$$

Gill dubs the first two components the "additive" part, and the second the "interactive" bit. Here, z will be the "response" variable, and we'll use ψ to introduce covariates. It's important to emphasize that any PDF that can be written in this way is a member of the exponential family; as we'll see, this encompasses a really broad range of interesting densities.

"Canonical" Forms

The "canonical" form of the distribution results when $q(z) = z$. This means that we can always "get to" the canonical form by transforming Z into Y according to $y = q(z)$. In canonical form, then, $Y = Z$ is our "response" variable; in such circumstances, $h(\psi)$ is referred to as the *natural parameter* of the distribution. We can write

$$\theta = h(\psi)$$

and recognize that, in some instances, $h(\psi) = \psi$ (so that $\theta = \psi$). This is known as the "link" function. Writing this allows us to collect terms from (2) into a simpler formulation:

$$f[y|\theta] = \exp[y\theta - b(\theta) + c(y)]. \tag{3}$$

[1] This presentation owes a good deal to Jeff Gill's exposition of GLMs at the Boston area JASA meeting, Feb. 28, 1998.

Here,

- $b(\theta)$ – sometimes called a "normalizing constant" – is a function solely of the parameter(s) θ,
- $c(y)$ is a function solely of y, the (potentially transformed) response variable, and
- $y\theta$ is a multiplicative term between the response and the parameter(s).

Some Familiar Exponential Family Examples

If all of this seems a bit abstract, it shouldn't. The key point to remember is that any probability distribution that can be written in the forms in (1) - (3) is a member of the exponential family. Consider three quick examples.

First, we know that if a variable Y follows a Poisson distribution, we can write its density as:

$$f(y|\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}. \tag{4}$$

Rearranging a bit, we can rewrite (4) as:

$$\begin{aligned} f(y|\lambda) &= \exp\left\{\ln\left[\frac{\exp(-\lambda)\lambda^y}{y!}\right]\right\} \\ &= \exp[y\ln(\lambda) - \lambda - \ln(y!)] \end{aligned} \tag{5}$$

Here, we have $\theta = \ln(\lambda)$, and (equivalently) $\lambda \equiv b(\theta) = \exp(\theta)$, while $c(y) = \ln(y!)$. The Poisson is thus an example of a one-parameter exponential family distribution.

Distributional parameters that are outside of the formulation in (1) - (3) are usually called *nuisance parameters*. They are typically not of central interest to researchers, though we still must treat them in the exponential family formulation. A common kind of nuisance parameter is a "scale parameter" – a parameter that defines the scale of the response variable (that is, it "stretches" or "shrinks" the axis of that variable). We can incorporate such a parameter (call it ϕ) through a slight generalization of (3):

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{6}$$

When no scale parameter is present, $a(\phi) = 1$ and (6) reverts to (3). We can illustrate such a distribution by considering our old friend, the normal distribution:

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right] \tag{7}$$

In the normal, σ^2 can be thought of as a scale parameter – it defines the "scale" (variance) of Y. We can rewrite (7) in exponential form as:

$$\begin{aligned} f(y|\mu,\sigma^2) &= \exp\left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right] \\ &= \exp\left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}y^2 + \frac{1}{2\sigma^2}2y\mu - \frac{1}{2\sigma^2}\mu^2\right] \\ &= \exp\left[\frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right] \\ &= \exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} + \frac{-1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right] \end{aligned} \tag{8}$$

Here, $\theta = \mu$, which means that:

- $y\theta = y\mu$,
- $b(\theta) = \frac{\mu^2}{2}$,
- $a(\phi) = \sigma^2$ is the scale (nuisance) parameter, and
- $c(y,\phi) = \frac{-1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)$.

The Normal distribution is thus a member of the exponential family, and is in canonical form as well. Other distributions that are canonical members of the family include the binomial (with p as the canonical parameter), the gamma, and the inverse gaussian. Non-canonical family members include the lognormal and Weibull distributions.

Little Red Likelihood

To estimate and make inferences about the parameters of interest, we can derive a general likelihood for the class of models expressable as (3). That (log-)likelihood is:

$$\begin{aligned} \ln L(\theta,\phi|y) &= \ln f(y|\theta,\phi) \\ &= \ln\left\{\exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]\right\} \\ &= \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi). \end{aligned} \tag{9}$$

This expresses the log-likelihood of the parameters in terms of the data. To obtain estimates, we maximize this function in the usual way: by taking its first derivative with respect to the parameter(s) of interest θ, setting it equal to zero, and solving. The former is:

$$\begin{aligned}\frac{\partial \ln L(\theta, \phi|y)}{\partial \theta} \equiv \mathbf{S} &= \frac{\partial}{\partial \theta}\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \\ &= \frac{y - \frac{\partial}{\partial \theta} b(\theta)}{a(\phi)}. \end{aligned} \tag{10}$$

In GLM-speak, the quantity in (10) – the partial derivative of the log-likelihood with respect to the parameter(s) of interest – is typically called the "score function." It has a number of important properties:

- $\mathbf{S}$ is a sufficient statistic for θ.
- $\mathrm{E}(\mathbf{S}) = 0$.
- $\mathrm{Var}(\mathbf{S}) \equiv \mathcal{I}(\theta) = \mathrm{E}[(\mathbf{S})^2|\theta]$, because $\mathrm{E}(\mathbf{S}) = 0$. This is known as the Fisher information (hence, $\mathcal{I}(\theta)$).

Finally, assuming general regularity conditions (which all exponential family distributions meet), it is the case that the conditional expectation of Y is just:

$$\mathrm{E}(Y) = \frac{\partial}{\partial \theta} b(\theta) \tag{11}$$

and

$$\mathrm{Var}(Y) = a(\phi)\frac{\partial^2}{\partial \theta^2} b(\theta) \tag{12}$$

(see, e.g., McCullagh and Nelder 1989, pp. 26-29 for a derivation). This means that – for exponential family distributions – we can get the mean of Y directly from $b(\theta)$, and the variance of Y from $b(\theta)$ and $a(\phi)$. So, for our Poisson example, (11) and (12) mean that:

$$\begin{aligned}\mathrm{E}(Y) &= \frac{\partial}{\partial \theta}\exp(\theta) \\ &= \exp(\theta)|_{\theta=\ln(\lambda)} \\ &= \lambda \end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(Y) &= 1 \times \frac{\partial^2}{\partial \theta^2}\exp(\theta)|_{\theta=\ln(\lambda)} \\ &= \exp[\ln(\lambda)] \\ &= \lambda \end{aligned}$$

which is exactly what we'd expect for the Poisson. Likewise, from our normal example above, (11) and (12) imply that:

$$\begin{aligned} \mathrm{E}(Y) &= \frac{\partial}{\partial \theta}\left(\frac{\theta^2}{2}\right) \\ &= \theta|_{\theta=\mu} \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(Y) &= \sigma^2 \times \frac{\partial^2}{\partial \theta^2}\left(\frac{\theta^2}{2}\right) \\ &= \sigma^2 \times \frac{\partial}{\partial \theta}\theta \\ &= \sigma^2 \end{aligned}$$

There are lots more interesting things about exponential family distributions, but we'll skip them in the interest of saving time.

Generalized Linear Models

What does all this have to do with regression-like models? To answer that, let's return to our old friend, the continuous linear-normal regression model:

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + u_i \tag{13}$$

where Y_i is the $N \times 1$ vector for the response/"dependent" variable, $\mathbf{X}_i$ is the $N \times k$ matrix of covariates, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters, and u_i is a $N \times 1$ vector of i.i.d. Normal errors with mean zero and constant variance σ^2. In this standard formulation, we usually think of $\mathbf{X}_i\boldsymbol{\beta}$ as the "systematic component" of Y_i, such that

$$\mathrm{E}(Y_i) \equiv \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} \tag{14}$$

In this setup, the variable Y is distributed Normally, with mean $\boldsymbol{\mu}_i$ and variance σ^2, and wed can estimate the parameters $\boldsymbol{\beta}$ (and σ^2) in the usual way.

The "Generalized" Part

We can generalize (14) by allowing the expected value of Y to vary with $\mathbf{X}\boldsymbol{\beta}$ according to a function $g(\cdot)$:

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta}. \tag{15}$$

$g(\cdot)$ is generally referred to as a "link function," because its purpose is to "link" the expected value of Y_i to $\mathbf{X}_i\boldsymbol{\beta}$ in a general, flexible way. We require $g(\cdot)$ to be continuously twice-differentiable (that is, "smooth") and monotonic in $\boldsymbol{\mu}$.

Importantly, note that we are not transforming the value of Y itself, but rather its expected value. In fact, in this formulation, $\mathbf{X}_i\boldsymbol{\beta}$ informs a function of $\mathrm{E}(Y_i)$, not $\mathrm{E}(Y_i)$ itself. In this way, $g[\mathrm{E}(Y_i)]$ is required to meet all the usual Gauss-Markov assumptions (linearity, homoscedasticity, normality, etc.) but $\mathrm{E}(Y_i)$ itself need not do so.

It is useful to write

$$\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta}$$

where $\boldsymbol{\eta}_i$ is often known as the "linear predictor" (or "index"); this means that we can rewrite (15) as:

$$\boldsymbol{\eta}_i = g(\boldsymbol{\mu}_i) \tag{16}$$

Of course, this also means that we can invert the function $g(\cdot)$ to express $\boldsymbol{\mu}_i$ as a function of $\mathbf{X}_i\boldsymbol{\beta}$:

$$\begin{aligned} \boldsymbol{\mu}_i &= g^{-1}(\boldsymbol{\eta}_i) \\ &= g^{-1}(\mathbf{X}_i\boldsymbol{\beta}) \end{aligned} \tag{17}$$

Pulling all this together, we can think of a GLM as having three key components:

1. The *random component* is the stochastic part of Y; it is distributed according to some exponential family distribution with mean

   $$\mathrm{E}(Y_i) = \boldsymbol{\mu}_i.$$

2. The *systematic component* is the covariates and their associated parameters:

   $$\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta}.$$

3. The *link* between the systematic and random parts of the model:

   $$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i$$

   and, equivalently,

   $$g^{-1}(\boldsymbol{\eta}_i) = \boldsymbol{\mu}_i.$$

The "generalized" part of GLM thus comes from two things: First, unlike the classical linear regression model, the distribution of $\boldsymbol{\mu}$ need not be Normal, and in fact can come from any member of the exponential family we discussed above. Second, the link function in (15) need not be an identity function, but instead can be any monotone, differentiable function.

Bringing In the Exponential Family

As we outlined above, we can think of distributions from the exponential family as having a canonical parameter $\boldsymbol{\theta}$. In a GLM, we link the systematic part of the model (that is, the covariates) to that parameter through the link function. Generically:

$$\begin{aligned} \boldsymbol{\theta}_i &= g(\boldsymbol{\mu}_i) \\ &= \boldsymbol{\eta}_i \\ &= \mathbf{X}_i\boldsymbol{\beta} \end{aligned}$$

which defines the expected value of Y. This means that, by setting $\boldsymbol{\theta}_i = \boldsymbol{\eta}_i$, we are allowing the transformed mean of the distribution to vary linearly in the covariates. In this light, one can imagine extending the same idea to any of the distributions in the exponential family, through the link function $g^{-1}(\cdot)$:

$$g^{-1}(\boldsymbol{\theta}_i) = \boldsymbol{\eta}_i \tag{18}$$

The link function that makes the linear predictor the same as the parameter $\boldsymbol{\theta}$ is known as the *canonical link function*. A useful aspect of the canonical link function is that it ensures that there is a sufficient statistic of rank k (that is, the number of parameters in $\boldsymbol{\beta}$).

Some Brief Examples

Linear-Normal

As a first example, consider data where Y is a continuous, unbounded, and normally-distributed variable. In such an instance, the natural conditional distribution to choose is the *Normal distribution* with mean μ and variance σ^2:

$$f(y|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$$

Similarly, because there is no need to restrict the effect of $\mathbf{X}$ on Y, the simplest (and most natural) link function is the *identity function*:

$$\boldsymbol{\mu}_i = \boldsymbol{\eta}_i.$$

This combination of link function and distributional family leads to a model with:

$$\begin{aligned} \boldsymbol{\mu}_i \equiv \boldsymbol{\theta}_i &= \boldsymbol{\eta}_i \\ Y_i &\sim N(\boldsymbol{\mu}_i, \sigma^2) \end{aligned}$$

that is, the standard linear-normal regression model.

Binary

Now suppose instead that our response variable Y is binary. The natural distributional family for a binary variable is the *Bernoulli* (that is, a binomial with $n = 1$):

$$f(y|\pi) = \pi^y (1-\pi)^{1-y}.$$

The canonical link function for the Bernoulli (and, in fact for the binomial in general) is the *logit*:

$$\boldsymbol{\theta}_i = \ln\left(\frac{\boldsymbol{\mu}_i}{1-\boldsymbol{\mu}_i}\right)$$

This restricts the possible outcomes to fall in the $[0, 1]$ interval, and thus restricts the impact of $\mathbf{X}$ as well. It yields a model with:

$$\begin{aligned} \boldsymbol{\mu}_i &= g^{-1}(\boldsymbol{\theta}_i) \\ &= \frac{\exp(\boldsymbol{\eta}_i)}{1+\exp(\boldsymbol{\eta}_i)} \\ Y_i &\sim \text{Bernoulli}(\boldsymbol{\mu}_i) \end{aligned}$$

that is, a standard logit model.

Count/Poisson

Finally, consider the case where Y_i is an event count. As we discussed some time ago, the natural distribution for a count of conditionally independent events is a Poisson:

$$f(y|\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}$$

The canonical parameter of the Poisson is $\boldsymbol{\theta} = \lambda$; because we require the counts to be strictly positive, the natural (and, as it happens, canonical) link function is the one which preserves the nonnegativity of λ, the log:

$$\ln(\lambda_i) = \boldsymbol{\eta}_i$$

which in turn yields a model with:

$$\begin{aligned} \boldsymbol{\mu}_i &= g^{-1}(\boldsymbol{\theta}_i) \\ &= \exp(\boldsymbol{\eta}_i) \\ Y_i &\sim \text{Poisson}(\lambda_i) \end{aligned}$$

that is, a Poisson model.

Table 1 lists several commonly-used GLMs, including their canonical link functions.

The biggest point here is that a large number of models – including the classical linear regression model, logit/probit/cloglog models for binary responses, and Poisson and negative binomial models for event counts – all fall under the GLM rubric. In addition, GLMs also encompass a number of models (such as the gamma and the inverse Gaussian, both for positive-continuous outcomes) that are not generally discussed in "MLE"-type courses. Finally, the GLM framework is a very general one, in that one can mix-and-match marginal distributions and link functions to fit specific data problems; we'll work through an example of this below.

Table 1: Common Flavors of GLMs

Distribution	Range of Y	Link(s)	Inverse Link
Normal	$(-\infty, \infty)$	Identity: $\boldsymbol{\theta} = \boldsymbol{\mu}$ (Canonical)	$\boldsymbol{\theta}$
Binomial	[0,1]	Logit: $\boldsymbol{\theta} = \ln\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right)$ (Canonical)	$\frac{\exp(\boldsymbol{\theta})}{1+\exp(\boldsymbol{\theta})}$
		Probit: $\boldsymbol{\theta} = \Phi^{-1}(\boldsymbol{\mu})$	$\Phi(\boldsymbol{\theta})$
		C-Log-Log: $\boldsymbol{\theta} = \ln[-\ln(1-\boldsymbol{\mu})]$	$1 - \exp[-\exp(\boldsymbol{\theta})]$
Poisson	$[0, \infty]$ (integers)	Log: $\boldsymbol{\theta} = \ln(\boldsymbol{\mu})$ (Canonical)	$\exp(\boldsymbol{\theta})$
Gamma	$(0, \infty)$	Reciprocal: $\boldsymbol{\theta} = -\frac{1}{\boldsymbol{\mu}}$ (Canonical)	$-\frac{1}{\boldsymbol{\theta}}$

Some Examples

As a practical matter, then, GLMs involve three steps:

1. Selecting a distribution $f(Y)$ for the stochastic component (from the family of exponential distributions) that best reflects the underlying data-generating process of Y.

2. Selecting a link function $g(\cdot)$ appropriate for the nature of the response Y.

3. Specifying the model – that is, choosing variables for inclusion into $\mathbf{X}$.

I won't go into detail about parameter estimation in the context of GLMs; suffice it to say that the more-or-less standard methods we're all familiar with (MLE, MCMC) can be used here as well. In addition, it is also possible to estimate GLMs using iteratively reweighted least squares (IRLS); see McCullagh and Nelder (1989, pp. 40-42). This has the advantage of being equivalent to MLE, but being computationally easier in many cases.

For now, we'll illustrate the equivalence of GLMs to many of the models we're already familiar with, using data from 7,161 decisions of the United States Supreme Court from 1953 to 1986 (that is, the entirety of the Warren and Burger Courts). The variables are:
and the primary variable of interest –

- `amici` – a count variable recording the number of *amicus curiae* ("friend of the Court") briefs filed in each case.
- `us` is an identifier code for the case in question (which can be ignored).
- `year` identifies the term in which the case was decided (minus 1900).
- `fedparty` is coded one in cases in which the federal government is a party (either petitioner or respondent) and zero otherwise; to the extent that the U.S. tends to win the cases in which it is involved, there may be correspondingly less room to influence the Court's decision (and groups may therefore be less likely to file briefs in such cases).
- `multlaw` is coded one in cases which involved multiple legal issues and zero otherwise; by their complexity, such cases present greater opportunities for groups to take an interest in the litigation, and therefore should have a positive influence on the number of amici filed.
- `conflict` is coded one in cases involving a conflict between two or more federal circuit courts; such cases are generally high-profile affairs, and therefore should be expected to lead to higher numbers of amicus filings.
- `constit` is a dummy variable, coded one in cases involving a constitutional issue. As with `conflict`, these tend to be higher-salience cases, and so should attract greater numbers of amici.

Summary statistics are:

```
. su

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       amici |      7161    .8419215    2.189712          0         39
          us |         0
        year |      7157    71.93014     9.20792         53         86
```

```
    fedparty |      7161    .3742494    .4839623          0          1
     multlaw |      7156    .1489659    .3560797          0          1
-------------+---------------------------------------------------------
    conflict |      7161    .0646558    .2459346          0          1
     constit |      7161    .2535959    .4350993          0          1
```

Replicating OLS, Logit, Probit, and Poisson

The results below illustrates how one can use the `glm` command in Stata to obtain results exactly like those we get using `regress`, `logit`, `probit`, and `poisson`, respectively. These are illustrated below. I won't go into a lot of detail on these, since there isn't much to say, except to note that:

1. Technically, `glm` yields exactly the same results as `regress`, `logit`, etc. only in those instances where the link function is the canonical one. In other instances – using a probit or c-log-log link with a binary response, for example – the standard errors are only asymptotically equivalent. As a practical matter, this is usually not a big issue; it certainly isn't in the example data, where the N is large.

2. There are also some special considerations to think about when using `glm` to estimate negative binomial models (not shown). Read more on this in the Stata manuals, or in one or more of the references below, if it is something you care about.

Linear Regression

```
. regress amici year fedparty multlaw conflict constit

      Source |       SS       df       MS              Number of obs =    7156
-------------+------------------------------           F(  5,  7150) =  106.72
       Model |  2383.90863     5  476.781725            Prob > F      =  0.0000
    Residual |  31943.5999  7150  4.46763635            R-squared     =  0.0694
-------------+------------------------------           Adj R-squared =  0.0688
       Total |  34327.5085  7155  4.79769511            Root MSE      =  2.1137
------------------------------------------------------------------------------
       amici |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0452468   .0027576    16.41   0.000     .0398411    .0506525
    fedparty |  -.1803145   .0537587    -3.35   0.001    -.2856975   -.0749315
     multlaw |   .7441085   .0731066    10.18   0.000      .600798    .8874191
    conflict |   .1720396   .1031404     1.67   0.095     -.030146    .3742253
     constit |   .4178814   .0604769     6.91   0.000     .2993287    .5364341
       _cons |  -2.572682   .2034335   -12.65   0.000    -2.971471   -2.173892
------------------------------------------------------------------------------

. glm amici year fedparty multlaw conflict constit, family(normal) link(identity)

Iteration 0:   log likelihood = -15506.686

Generalized linear models                         No. of obs      =      7156
Optimization     : ML                             Residual df     =      7150
                                                  Scale parameter =  4.467636
Deviance         =   31943.5999                   (1/df) Deviance =  4.467636
Pearson          =   31943.5999                   (1/df) Pearson  =  4.467636
Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  4.335575
Log likelihood   = -15506.68613                   BIC             =  -31517.7
------------------------------------------------------------------------------
             |                 OIM
       amici |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0452468   .0027576    16.41   0.000      .039842    .0506516
    fedparty |  -.1803145   .0537587    -3.35   0.001    -.2856796   -.0749494
     multlaw |   .7441085   .0731066    10.18   0.000     .6008223    .8873948
    conflict |   .1720396   .1031404     1.67   0.095    -.0301118     .374191
     constit |   .4178814   .0604769     6.91   0.000     .2993488    .5364141
       _cons |  -2.572682   .2034335   -12.65   0.000    -2.971404   -2.173959
------------------------------------------------------------------------------
```

Logit

```
. logit multlaw year fedparty conflict constit

Logistic regression                                     Number of obs   =       7156
                                                        LR chi2(4)      =     521.28
                                                        Prob > chi2     =     0.0000
Log likelihood =  -2751.402                             Pseudo R2       =     0.0865
------------------------------------------------------------------------------
     multlaw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0014409   .0038687    -0.37   0.710    -.0090234    .0061416
    fedparty |   .7692011   .0733846    10.48   0.000     .6253699    .9130322
    conflict |  -.3407245   .1594579    -2.14   0.033    -.6532563   -.0281927
     constit |   1.563187   .0728051    21.47   0.000     1.420492    1.705883
       _cons |   -2.48845   .2851855    -8.73   0.000    -3.047403   -1.929496
------------------------------------------------------------------------------

. glm multlaw year fedparty conflict constit, family(binomial) link(logit)

Generalized linear models                               No. of obs      =      7156
Optimization     : ML                                   Residual df     =      7151
                                                        Scale parameter =         1
Deviance         =  5502.803907                         (1/df) Deviance =  .7695153
Pearson          =  7436.082429                         (1/df) Pearson  =  1.039866

Variance function: V(u) = u*(1-u/1)                     [Binomial]
Link function    : g(u) = ln(u/(1-u))                   [Logit]

                                                        AIC             =  .7703751
Log likelihood   = -2751.401954                         BIC             = -57967.37
------------------------------------------------------------------------------
             |                 OIM
     multlaw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0014409   .0038687    -0.37   0.710    -.0090234    .0061416
    fedparty |   .7692011   .0733846    10.48   0.000     .6253699    .9130322
    conflict |  -.3407245   .1594579    -2.14   0.033    -.6532563   -.0281927
     constit |   1.563187   .0728051    21.47   0.000     1.420492    1.705883
       _cons |   -2.48845   .2851855    -8.73   0.000    -3.047403   -1.929496
------------------------------------------------------------------------------
```

Probit

```
. probit multlaw year fedparty conflict constit

Probit regression                                 Number of obs   =       7156
                                                  LR chi2(4)      =     507.62
                                                  Prob > chi2     =     0.0000
Log likelihood = -2758.2281                       Pseudo R2       =     0.0843
------------------------------------------------------------------------------
     multlaw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0012717   .0020975    -0.61   0.544    -.0053829    .0028394
    fedparty |   .3899633   .0398029     9.80   0.000     .3119511    .4679755
    conflict |  -.1650351   .0826823    -2.00   0.046    -.3270895   -.0029807
     constit |   .8563234    .040428    21.18   0.000     .7770861    .9355608
       _cons |  -1.380112   .1540479    -8.96   0.000     -1.68204   -1.078184
------------------------------------------------------------------------------

. glm multlaw year fedparty conflict constit, family(binomial) link(probit)

Generalized linear models                         No. of obs      =      7156
Optimization     : ML                             Residual df     =      7151
                                                  Scale parameter =         1
Deviance         =  5516.456215                   (1/df) Deviance =  .7714244
Pearson          =  7379.781144                   (1/df) Pearson  =  1.031993

Variance function: V(u) = u*(1-u/1)               [Binomial]
Link function    : g(u) = invnorm(u)              [Probit]

                                                  AIC             =  .7722829
Log likelihood   = -2758.228108                   BIC             = -57953.72
------------------------------------------------------------------------------
             |                 OIM
     multlaw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0012717   .0020975    -0.61   0.544    -.0053829    .0028394
    fedparty |   .3899633   .0398029     9.80   0.000     .3119511    .4679755
    conflict |  -.1650351   .0826823    -2.00   0.046    -.3270895   -.0029807
     constit |   .8563234    .040428    21.18   0.000     .7770861    .9355608
       _cons |  -1.380112    .154048    -8.96   0.000     -1.68204   -1.078183
------------------------------------------------------------------------------
```

Poisson

```
. poisson amici year fedparty multlaw conflict constit

Poisson regression                                Number of obs   =       7156
                                                  LR chi2(5)      =    2851.28
                                                  Prob > chi2     =     0.0000
Log likelihood = -11221.591                       Pseudo R2       =     0.1127
------------------------------------------------------------------------------
       amici |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0630961   .0016768    37.63   0.000     .0598097    .0663826
    fedparty |  -.2860027   .0293785    -9.74   0.000    -.3435835   -.2284219
     multlaw |   .6871808   .0307847    22.32   0.000      .626844    .7475177
    conflict |   .1822242   .0504315     3.61   0.000     .0833803    .2810681
     constit |   .4335846   .0279461    15.52   0.000     .3788112     .488358
       _cons |  -5.077352   .1305304   -38.90   0.000    -5.333187   -4.821517
------------------------------------------------------------------------------

. glm amici year fedparty multlaw conflict constit, family(poisson) link(log)

Generalized linear models                         No. of obs      =       7156
Optimization     : ML                             Residual df     =       7150
                                                  Scale parameter =          1
Deviance         =  17031.76219                   (1/df) Deviance =   2.382065
Pearson          =  33685.24336                   (1/df) Pearson  =   4.711223

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   3.137952
Log likelihood   = -11221.59072                   BIC             = -46429.54
------------------------------------------------------------------------------
             |                 OIM
       amici |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0630961   .0016768    37.63   0.000     .0598097    .0663826
    fedparty |  -.2860027   .0293785    -9.74   0.000    -.3435835   -.2284219
     multlaw |   .6871808   .0307847    22.32   0.000      .626844    .7475177
    conflict |   .1822242   .0504315     3.61   0.000     .0833803    .2810681
     constit |   .4335846   .0279461    15.52   0.000     .3788112     .488358
       _cons |  -5.077352   .1305304   -38.90   0.000    -5.333187   -4.821517
------------------------------------------------------------------------------
```

An Extension: Models of Supreme Court Coalition Size

GLMs can also do things that more "standard" MLE-style models heretofore discussed cannot do. To take an example, consider the question of the size of a majority opinion coalition on the U.S. Supreme Court. We know that:

- To constitute a majority, the opinion coalition must be made up of a number of justices equal to of greater than half of the justices who heard the case. However,
- Not every case is heard and voted on by every justice – sometimes there are vacancies, sometimes justices recuse themselves for various reasons, and sometimes both occur.

As a practical matter, then, majority coalition sizes range from three (in 3-2 decisions, say, where four justices are absent or recuse) to nine (in 9-0 unanimous cases). But the size of the majority coalition is always at least partially determined by the number of justices voting in the case.

Suppose we define N_i as the number of justices voting in case i, and denote the size of the majority coalition in that case as M_i. Then it makes some sense[2] to think of M_i as a binomial function:

$$M_i \sim \binom{N_i}{p_i} p_i^{M_i} (1 - p_i)^{N_i - M_i}$$

We can estimate a model of a variable like M_i using a GLM framework. In particular, we can estimate a model of M_i where N_i is allowed to vary across i, and where p_i – the probability of a marginal justice joining the majority coalition – is allowed to vary systematically with covariates $\mathbf{X}_i$. By choosing the binomial distribution from the exponential family, and a logit link (the canonical link for the binomial distribution, and a natural choice, given that p_i is a probability), we get a model which is:

$$\begin{aligned} \mathrm{E}(M_i) &= \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})} \\ Y_i &\sim \mathrm{Binomial}(M_i, N_i) \end{aligned}$$

We'll estimate this GLM on data on the U.S. Supreme Court between 1953 and 2005 (a bit larger dataset than the one for the examples above), focusing on three covariates:

- `term`, the term in which the decision was handed down,
- `alt_prec`, a dummy variable indicating whether (=1) or not (=0) the decision in question altered existing Supreme Court precedent, and

[2]To be brutally honest, this is not completely right. What we'd *really* want to do is to have a binomial with a lower limit at the integer value of $0.5N_i$, since we're only looking at majority coalitions. But, this will do for pedagogical purposes.

- `unconst`, an indicator of whether (=1) or not (=0) the decision formally declared a federal, state, or local statute to be unconstitutional.

Summary statistics on the data look like this:

```
. su nmajority nvoting term alt_prec unconst

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   nmajority |     12491    7.045633    1.474143         3          9
     nvoting |     12580    8.585453     .754199         4          9
        term |     12583    1977.332    13.03965      1953       2005
    alt_prec |     12583    .0168481    .1287074         0          1
     unconst |     12583    .0665978     .249334         0          1
```

The results of the `glm` estimation are:

```
. glm nmajority term alt_prec unconst, family(binomial nvoting) link(logit) irls

Generalized linear models                         No. of obs      =     12491
Optimization     : MQL Fisher scoring             Residual df     =     12487
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  25766.28517                   (1/df) Deviance =  2.063449
Pearson          =  20745.56508                   (1/df) Pearson  =  1.661373

Variance function: V(u) = u*(1-u/nvoting)         [Binomial]
Link function    : g(u) = ln(u/(nvoting-u))       [Logit]

                                                  BIC             = -92020.63
------------------------------------------------------------------------------
             |                 EIM
   nmajority |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        term |  -.0014731   .0006069    -2.43   0.015    -.0026626   -.0002837
    alt_prec |  -.2685583   .0560735    -4.79   0.000    -.3784603   -.1586563
     unconst |  -.2798177   .0294079    -9.52   0.000    -.3374563   -.2221792
       _cons |   4.457874   1.200241     3.71   0.000     2.105444    6.810304
------------------------------------------------------------------------------
```

Note that we specify the number of "trials" in the binomial in the `family()` option to the `glm` command; here, that is equal to the variable `nvoting`. The results suggest that

- The sizes of majority coalitions declined (on average) over the past 50 years,

- Both precedent-altering decisions and those declaring statutes unconstitutional garnered smaller average majority coalitions than decisions that did not do so. This is unsurprising, given that such decisions are more likely to be controversial, and therefore less likely to command large majorities or a unanimous Court.

Of course, the nicest thing about this model is that the underlying statistical model is an especially nice "fit" to the data-generating process.

Software

As the examples illustrate, one can do GLMs all day long using Stata. Its `glm` command is very flexible; options are:

- `family`, which includes
  - `gaussian` (that is, normal),
  - `igaussian` (inverse Gaussian),
  - `binomial` (Bernoulli/binomial),
  - `poisson` (Poisson, duh...),
  - `nbinomial` (negative binomial),
  - `gamma` (gamma).
- `link`, which have:
  - `identity` ($\boldsymbol{\mu} = \boldsymbol{\eta}$)
  - `log` ($\ln(\boldsymbol{\mu}) = \boldsymbol{\eta}$),
  - `logit` ($\ln\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right) = \boldsymbol{\eta}$)
  - `probit` ($\Phi^{-1}(\boldsymbol{\mu}) = \boldsymbol{\eta}$)
  - `cloglog` ($\ln[-\ln(1-\boldsymbol{\mu})] = \boldsymbol{\eta}$)
  - `power` ($\boldsymbol{\mu}^{\ell} = \boldsymbol{\eta}$, with $\ell \in \{...-2,-1,0,1,2,...\}$)
  - `opower` ["odds power," $\frac{\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right)^{\ell}-1}{\ell}$, with ℓ as in `power`],
  - `nbinomial` (negative binomial, $\ln\left(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+k}\right) = \boldsymbol{\eta}$, with $k = 1$ or user-defined),
  - `loglog` ($-\ln[-\ln(\boldsymbol{\mu})] = \boldsymbol{\eta}$)
  - `logc` ("log-complement," $\ln(1-\boldsymbol{\mu}) = \boldsymbol{\eta}$).

One can also specify an `exposure()` or `offset()` variable, where appropriate, "robust" variance-covariance estimates, the method of maximization (IRLS or MLE), and other options specific to the distributions in question. Finally, specifying the `eform` option tells Stata to report $\exp(\hat{\boldsymbol{\beta}})$ rather than $\hat{\boldsymbol{\beta}}$; this is handy when the exponentiated coefficients take on useful forms (as is the case when the model is a logit, or a Poisson or negative binomial). The Stata help for `glm` helpfully notes the following:

```
If you specify both family() and link(), note that not all combinations make
sense.  You may choose from the following combinations:

          | id  log  logit  probit  clog  pow  opower  nbinomial  loglog  logc
----------+-------------------------------------------------------------------
Gaussian  |  x   x                          x
inv. Gau. |  x   x                          x
binomial  |  x   x     x      x       x     x     x                  x      x
Poisson   |  x   x                          x
neg. bin. |  x   x                          x             x
gamma     |  x   x                          x
```

Post-estimation, one can do all the "usual" things, including `estat`, `predict`, `mfx`, `test`, `testnl`, etc. Many of these take on different forms as a function of the distributional family chosen, so consult your manuals.

Beyond Stata , there are a host of other programs out there that will estimate GLMs. The major ones are:

- R (with the `glm` package),
- SAS (with `proc glm`)
- SPSS (with the `glm` library),
- GLIM, Genstat, LispStat, and many others are free-standing packages that will estimate GLMs.

References

One could easily teach an entire course on GLMs – and, in fact, most statistics departments do just that. There are, as a result, a large number of books on the subject. Here are a sample, in no particular order, with my own thoughts on them:

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd Ed. London: Chapman & Hall.

This is the "bible" of GLMs. It's expensive, a bit dated, and the examples aren't exactly compelling to most social scientists (e.g., the length of time blood takes to clot, or growth rates of *Neurospora crassa* fungi), but it's still a great reference.

Hardin, James W., and Joseph W. Hilbe. 2007. *Generalized Linear Models and Extensions*, 2nd Ed. College Station, TX: Stata Press.

If you're going to use Stata to do GLMs, this might be the one to have. It's recent, written in a relatively clear notation, and has profuse examples. The down-side is its cost, and its weddedness to Stata for examples, etc.

Dobson, Annette J. 2001. *An Introduction to Generalized Linear Models*, 2nd Ed. London: Chapman & Hall.

A relatively concise, not-too-technical introduction to GLMs. It's also more-or-less current, includes chapters on a few related methods (like HLMs), and – best of all – can be had in paperback for relatively little money.

Gill, Jeff. 2005. *Generalized Linear Models: A Unified Approach*. Thousand Oaks, CA: Sage.

This is the "little green book" on GLMs. For a Sage monograph, it's nicely done, with good coverage and lots of nice examples (all of which are available on the author's website). And, as always for the Sage books, the price is right.

Faraway, Julian J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects, and Nonparametric Regression Models*. London: Chapman & Hall.

This is a terrific, compact book that covers a wide range of topics including (but also well beyond) GLMs. It's also fully integrated with the R statistical language, which makes it a great reference as well. The downsides are it doesn't muck around (it's remarkably terse for the range of topics covered – a fact some will surely find a benefit rather than a drawback) and, like many statistics texts, it's very short on social science examples.