GLMs; definition and derivation

Ben Bolker

September 21, 2018

Introduction

Definition:

- exponential family conditional distribution (all we will really use in fitting is the *variance function* $V(\mu)$: makes *quasi-likelihood models* possible)
- linear model $\boldsymbol{\eta}$ (*linear predictor*) $= \mathbf{X}\boldsymbol{\beta}$
- smooth, monotonic link function $\eta = g(\mu)$

Before we used

$$f(y;\theta,\phi) = \exp[(a(y)b(\theta) + c(\theta))/f(\phi) + d(y,\phi)]$$

but let's say without loss of generality (putting the distribution into **canonical form**):
$\{a(y) \mapsto y, b(\theta) \mapsto \theta, c(\theta) \mapsto -b(\theta), f(\phi) \mapsto \phi, d(y,\phi) \mapsto c(y,\phi)\}$ [1]:

$$\boxed{\ell = (y\theta - b(\theta))/(\phi/w) + c(y,\phi)}$$

where y=data, θ=*location parameter*, ϕ= *dispersion parameter* (*scale parameter*). Will mostly ignore the *a priori* weights w in what follows.

The **canonical link function** ($\mu \to \eta$) is g such that $g^{-1} = b$.

Example: Poisson distribution: use $\theta = \log(\lambda)$.

$$\begin{aligned} \ell(y,\lambda) &= y\log(\lambda) - \lambda - \log(y!) \\ \theta &= \log(\lambda) \\ \ell(y,\theta) &= y\theta - \exp(\theta) - \log(y!) \end{aligned} \tag{1}$$

so $b = \exp$; $\phi = 1$; $c = -\log(y!)$. Canonical link is $\log(\mu) = \theta$.

[1] McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall; and

Useful facts

- *The* score function $\boldsymbol{u} = \frac{\partial \ell}{\partial \theta}$ *has expected value zero.*
- Therefore for exponential family:

$$\begin{aligned} E((y - b'(\theta))/\phi) &= 0 \\ (\mu - b'(\theta))/\phi &= 0 \\ \mu &= b'(\theta) \end{aligned} \tag{2}$$

(Check against Poisson.)

- Mean depends *only* on $b'(\theta)$.

  Variance calculation:

- For log-likelihood ℓ,

$$E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = -E\left(\frac{\partial \ell}{\partial \theta}\right)^2 \tag{3}$$

- Therefore for exponential family:

$$\begin{aligned} E\left(\frac{b''(\theta)}{\phi}\right) &= -E\left(\frac{Y - b'(\theta)}{\phi}\right)^2 \\ \frac{b''(\theta)}{\phi} &= -\frac{\mathrm{var}(Y)}{\phi^2} \\ \mathrm{var}(Y) &= b''(\theta)\phi = \frac{\partial \mu}{\partial \theta}\phi \equiv V(\mu)\phi \end{aligned} \tag{4}$$

- Check against Poisson.
- Variance depends *only* on $b''(\theta)$ and ϕ.

Iteratively reweighted least squares

Procedure

Likelihood equations

- compute **adjusted dependent variate**:

$$Z_0 = \hat{\eta}_0 + (Y - \hat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)_0$$

  (note: $\frac{d\eta}{d\mu} = \frac{d\eta}{dg(\eta)} = 1/g'(\eta)$: translate from raw to linear predictor scale)

- compute **weights**

$$W_0^{-1} = \left(\frac{d\eta}{d\mu}\right)_0^2 V(\hat{\mu}_0)$$

  (translate variance from raw to linear predictor scale). This is the inverse variance of Z_0.

- regress z_0 on the covariates with weights W_0 to get new $\boldsymbol{\beta}$ estimates ($\rightarrow$ new $\boldsymbol{\eta}$, $\boldsymbol{\mu}$, $V(\mu)$ ...)

Tricky bits: starting values, non-convergence, etc.. (We will worry about these later!)

Justification

Reminders:

- Maximum likelihood estimation (consistency; asymptotic Normality; asymptotic efficiency; "when it can do the job, it's rarely the best tool for the job but it's rarely much worse than the best" (S. Ellner); flexibility)
- multidimensional Newton-Raphson estimation: iterate solution of $\boldsymbol{H\beta} = \boldsymbol{u}$ where $\boldsymbol{H}$ is the negative of the *Hessian* (second-derivative matrix of ℓ wrt $\boldsymbol{\beta}$), $\boldsymbol{u}$ is the *gradient* or *score* vector (derivatives of ℓ wrt $\boldsymbol{\beta}$)

Maximum likelihood equations Remember $\ell = \sum_i w_i \left((y_i\theta_i - b(\theta_i))/\phi + c(y, \phi) \right)$. Ignore the last term because it's independent of θ.

Partial Decompose $\frac{\partial \ell}{\partial \beta_j}$ into

$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell}{\partial \theta} \cdot \frac{\partial \theta}{\partial \mu} \cdot \frac{\partial \mu}{\partial \eta} \cdot \frac{\partial \eta}{\partial \beta_j} \tag{5}$$

- $\frac{\partial \ell}{\partial \theta}$: effect of θ on log-likelihood, $(Y - \mu)/\phi$.
- $\frac{\partial \theta}{\partial \mu}$: effect of mean on θ. $d\mu/d\theta = d(b')/d\theta = b'' = V(\mu)$, so this term is $1/V$.
- $\frac{\partial \mu}{\partial \eta}$: dependence of mean on η (this is just the inverse-link function)
- $\frac{\partial \eta}{\partial \beta_j}$: the linear predictor $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, so this is just x_j.

So we get

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= \frac{(Y - \mu)}{\phi} \cdot \frac{1}{V} \cdot \frac{d\mu}{d\eta} \cdot x_j \\ &= \frac{1}{\phi} W (Y - \mu) \frac{d\eta}{d\mu} x_j \end{aligned} \tag{6}$$

This gives us a likelihood (score) equation

$$\sum u = \sum W(y - \mu) \frac{d\eta}{d\mu} x_j = 0 \tag{7}$$

(remember $W = (d\mu/d\eta)^2/V$) (this expression ignores *a priori* weights w on the variables, which we use in binomial regression). We can also express the vector as $W \frac{d\eta}{d\mu} \mathbf{X}^T (\mathbf{y} - \boldsymbol{\mu})$.

Scoring method Going back to finding solutions of the score equation: what is $\boldsymbol{H}$? (We're going to flip the sign of the score $\boldsymbol{u}$ now ...)

$$\begin{aligned}\boldsymbol{H}_{rs} &= -\frac{\partial u_r}{\partial \beta_s} \\ &= \sum \left[(Y-\mu) \frac{\partial}{\partial \beta_s} \left(W \frac{d\eta}{d\mu} x_r \right) \right. \\ &\qquad \left. + W \frac{d\eta}{d\mu} x_r \frac{\partial}{\partial \beta_s} (Y - \mu) \right] \end{aligned} \tag{8}$$

The first term disappears if we take the *expectation* of the Hessian (*Fisher scoring*) *or* if we use a canonical link. (Explanation of the latter: $W d\eta/d\mu$ is constant in this case. For a canonical link $\eta = \theta$, so $d\mu/d\eta = db'(\theta)/d\theta = b''(\theta)$. Thus $W d\eta/d\mu = 1/V (d\mu/d\eta)^2 d\eta/d\mu = 1/V d\mu/d\eta = 1/b''(\theta) \cdot b''(\theta) = 1$.) (Most GLM software just uses Fisher scoring regardless of whether the link is canonical or non-canonical.)

The second term is

$$\sum W \frac{d\eta}{d\mu} x_r \frac{\partial \mu}{\partial \beta_s} = \sum W x_r x_s$$

(the sum is over observations) or $\mathbf{X}^T \boldsymbol{W} \mathbf{X}$ (where $\boldsymbol{W} = \text{diag}(W)$)

Then we have (ignoring ϕ)

$$\begin{aligned} \boldsymbol{H\beta}^* &= \boldsymbol{H\beta} + \boldsymbol{u} \\ \mathbf{X}^T \boldsymbol{W} \mathbf{X} \boldsymbol{\beta}^* &= \mathbf{X}^T \boldsymbol{W} \mathbf{X} \boldsymbol{\beta} + \boldsymbol{u} \\ &= \mathbf{X}^T \boldsymbol{W} (\mathbf{X}\boldsymbol{\beta}) + \mathbf{X}^T \boldsymbol{W} (y - \mu) \frac{d\eta}{d\mu} \\ &= \mathbf{X}^T \boldsymbol{W} \boldsymbol{\eta} + \mathbf{X}^T \boldsymbol{W} (y-\mu) \frac{d\eta}{d\mu} \\ \mathbf{X}^T \boldsymbol{W} \mathbf{X} \boldsymbol{\beta}^* &= \mathbf{X}^T \boldsymbol{W} \boldsymbol{z} \end{aligned} \tag{9}$$

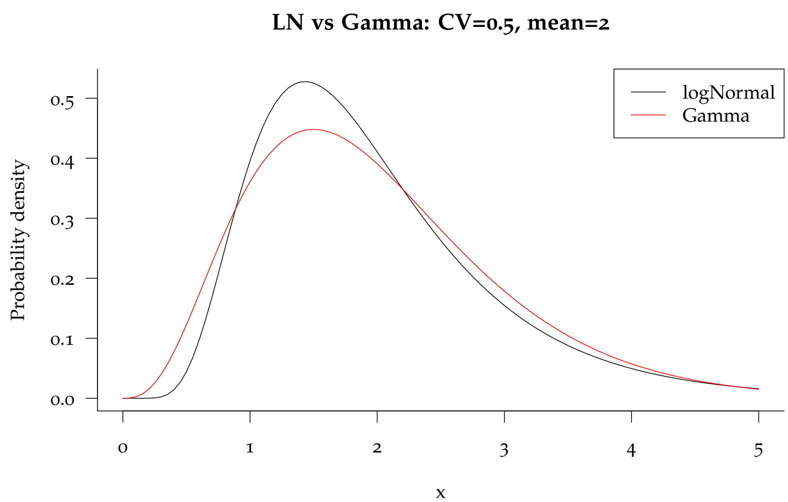
This is the same form as a weighted regression ... so we can use whatever linear algebra tools we already know for doing linear regression (QR/Cholesky decomposition, etc.)

Other sources

- (McCullagh and Nelder, 1989) is really the derivation of IRLS I like best, although I supplemented it at the end with (Dobson and Barnett, 2008).
- (Myers et al., 2010) has information about Newton-Raphson with non-canonical links.
- more details on fitting: (Marschner, 2011), interesting blog posts by Andrew Gelman, John Mount

Choice of distribution As previously discussed, choice of distribution should *usually* be dictated by data (e.g. binary data=binomial, counts of a maximum possible value=binomial, counts=Poisson ...) however, there is sometimes some wiggle room (Poisson with offset vs. binomial for rare counts; Gamma vs log-Normal for positive data). Then:

- Analytical convenience
- Computational convenience (e.g. log-Normal > Gamma; Poisson > binomial?)
- Interpretability (e.g. Gamma for multi-hit model)
- Culture (follow the herd)
- Goodness of fit (if it really makes a difference)



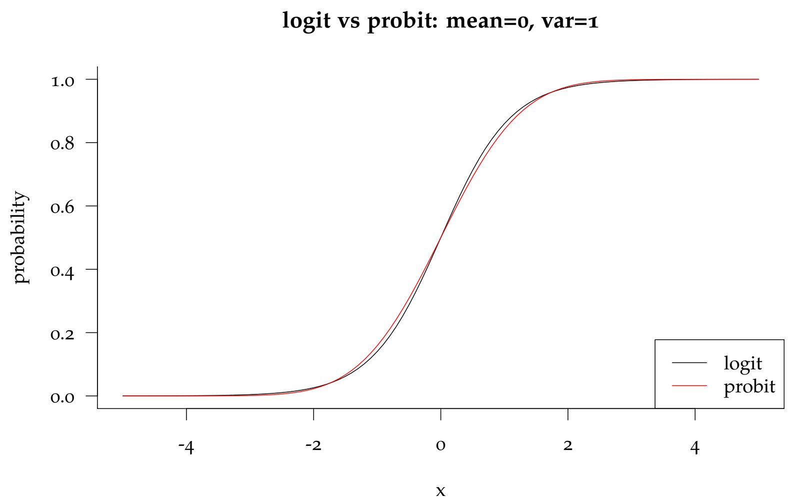
(*Note*: I cheated a little bit. The differences are smaller for smaller CVs/larger shape parameters ...)

Choice of link function More or less the same reasons, e.g.:

- analytical: canonical link best (logistic > probit: $g = \Phi^{-1}$)
- computational convenience: logistic > probit
- interpretability:
  - probit > logistic (latent variable model)
  - complementary log-log works well with variable exposure models

  - log link: proportional effects (e.g. multiplicative risk models in predator-prey settings)
  - logit link: proportional effects on odds

- culture: depends (probit in toxicology, logit in epidemiology ...)
- restriction of parameter space (log > inverse for Gamma models, because then range of g^{-1} is $(0, \infty)$)
- Goodness of fit: probit *very* close to logit



References

Dobson, A. J. and A. Barnett (2008, May). *An Introduction to Generalized Linear Models, Third Edition* (3 ed.). Chapman and Hall/CRC.

Marschner, I. C. (2011, December). glm2: Fitting generalized linear models with convergence problems. *The R Journal* 3(2), 12–15.

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

Myers, R. H., D. C. Montgomery, G. G. Vining, and T. J. Robinson (2010). Appendix A.6: Computational details for GLMs for a non-canonical link. In *Generalized Linear Models*, pp. 481–483. John Wiley & Sons, Inc.