Notes for week 3 (part 2)
Ben Bolker
September 25, 2013

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 $\eta$ (linear predictor) = $X\beta$
• smooth, monotonic link function $\eta = g(\mu)$

Reminder about the exponential family (notation from [3]):

$$\ell = \frac{(Y\theta - b(\theta))}{a(\phi)} + c(Y, \phi)$$

where $Y=$ data, $\theta=$ location parameter, $\phi=$ dispersion parameter (scale parameter). (This is written slightly differently from 1.)

May be useful to keep the definitions the Poisson distribution in mind to check against:

$$\ell(Y, \theta, \phi) = Y(\log \theta) - \exp(\log \theta) - \log(Y!)$$

so $b = \exp(\theta); a=$ identity; $\phi = 1; c = -\log(Y!)$

Useful facts

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0$$

$$E((Y - b'(\theta))/a(\phi)) = 0$$

$$\mu - b'(\theta)/a(\phi) = 0$$

$$\mu = b'(\theta)$$

• Check against Poisson.
• Mean depends only on $b'(\theta)$.
\[
\begin{align*}
E \left( \frac{\partial^2 \ell}{\partial \theta^2} \right) &= -E \left( \frac{\partial \ell}{\partial \theta} \right)^2 \\
E \left( \frac{b''(\theta)}{a(\phi)} \right) &= -E \left( \frac{Y - b'(\theta)}{a(\phi)} \right)^2 \\
\frac{b''(\theta)}{a(\phi)} &= -\frac{\text{var}(Y)}{a^2(\phi)} \\
\text{var}(Y) &= b''(\theta)a(\phi) = \frac{\partial \mu}{\partial \theta} a(\phi) \equiv V(\mu)a(\phi) \\
\text{(3)}
\end{align*}
\]

- Check against Poisson.
- Variance depends only on \( b''(\theta) \) and \( a(\phi) \).

Usually have \( a(\phi) = \phi/w \) where \( w \) are weights.

*Canonical link* uses \( g^{-1} = b \).

**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 larger for lower CV values …)

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

- compute **weights**

\[ W_0^{-1} = \left( \frac{d\eta}{d\mu} \right)_0 \sqrt{V(\hat{\mu}_0)} \]

(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 \( \beta \) estimates (\( \rightarrow \) new \( \eta, \mu, V(\mu) \ldots \))

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 $Adb = u$ where $A$ is the negative of the Hessian (second-derivative matrix of $\ell$ wrt $\beta$), $u$ is the gradient or score vector (derivatives of $\ell$ wrt $\beta$)

**Maximum likelihood equations**

Remember $\ell = (Y\theta - b(\theta))/a(\phi) + c(Y, \phi)$.

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

$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell}{\partial \theta} \cdot \frac{\partial \theta}{\partial \mu} \cdot \frac{\partial \mu}{\partial \eta} \cdot \frac{\partial \eta}{\partial \beta_j} \tag{4}$$

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

So we get

$$\frac{\partial \ell}{\partial \beta_j} = \frac{(Y - \mu)}{a(\phi)} \cdot \frac{1}{V} \cdot \frac{d\mu}{d\eta} \cdot x_j$$

$$= \frac{W}{a(\phi)} (Y - \mu) \frac{d\eta}{d\mu} x_j \tag{5}$$

Ignoring weights, this gives us a likelihood (score) equation

$$\sum u = \sum W(y - \mu) \frac{d\eta}{d\mu} x_j = 0 \tag{6}$$

**Scoring method**

Going back to finding solutions of the score equation: what is $A$?

$$A_{rs} = -\frac{\partial u_r}{\partial \beta_s}$$

$$= \sum \left[(Y - \mu) \frac{\partial}{\partial \beta_s} \left(W \frac{d\eta}{d\mu} x_r\right) + W \frac{d\eta}{d\mu} x_r \frac{\partial}{\partial \beta_s} (Y - \mu)\right] \tag{7}$$
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: $Wd\eta/d\mu$ is constant in this case. For a canonical link $\eta = \theta$, so $d\mu/d\eta = db'(\theta)/d\theta = b''(\theta)$. Thus $Wd\eta/d\mu = 1/V(d\mu/d\eta)^2d\eta/d\mu = 1/Vd\mu/d\eta = 1/b''(\theta) \cdot b''(\theta) = 1$.) (Most GLM software just uses Fisher scoring regardless of whether the link is canonical or non-canonical.)

The second term is

$$\sum Wx^{r}x_{s} = \sum Wx_{r}x_{s}$$

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

Then we have

$$Ab^{*} = Ab + u$$

$$X^{T}WXb^{*} = X^{T}WXb + u$$

$$= X^{T}W(Xb) + X^{T}(y - \mu)\frac{d\eta}{d\mu}$$

(8)

$$= X^{T}W\eta + X^{T}W(y - \mu)\frac{d\eta}{d\mu}$$

$$= X^{T}Wz$$

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


References

