The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1863-10

Advanced School and Conference on Statistics and Applied Probability in Life Sciences

24 September - 12 October, 2007

Generalized linear models II

Peter McCullagh

Department of Statistics
University of Chicago
Chicago IL 60637, USA

Generalized linear models II

Peter McCullagh

Department of Statistics
University of Chicago

Trieste, October 2007

Outline

Generalized linear models

Framework for non-Gaussian regression models:
- Retain exchangeability assumption
- Retain independence assumption

Extensions:
- Distribution: exponential family
  $E(Y_i) = \mu_i, \quad \text{var}(Y_i) = \sigma^2 V(\mu_i)$
- Non-linearity:
  $\eta_i = g(\mu_i), \quad \eta = X\beta$

Linear part: $\eta = X\beta$ (specified by a model formula)
Link function: $\eta_i = g(\mu_i)$
Distributional part: Exponential family

Parameters: regression coefficients β and variance $\sigma^2 > 0$.

Exponential families

Start off with baseline density $f_0(y)$ on $\mathcal{R}$
Weighted distribution with density proportional to $e^{\theta y} f_0(y)$
Normalization constant is $M_0(\theta) = \int e^{\theta y} f_0(y)\, dy$
Weighted density is

$$\begin{aligned} f_\theta(y) &= e^{\theta y} f_0(y)/M(\theta) \\ &= e^{\theta y - K(\theta)} f_0(y) \end{aligned}$$

$M_0(\theta)$ is the moment generating function of f_0
$K_0(\theta) = \log M_0(\theta)$ is the cumulant generating function.

Moments and cumulants

Moment generating function of f_θ is

$$\begin{aligned} M_\theta(t) &= \int e^{ty} f_\theta(y)\, dy \\ &= \int e^{(t+\theta)y} f_0(y)\, dy / M_0(\theta) = \frac{M_0(\theta+t)}{M_0(\theta)} \end{aligned}$$

for $\theta \in \Theta$

Cumulant generating function of f_θ is

$$K_\theta(t) = \log\left(\frac{M_0(\theta+t)}{M_0(\theta)}\right) = K_0(\theta+t) - K_0(\theta)$$

$\Rightarrow$ rth cumulant of f_θ is the rth derivative of K_0 at θ.

Moments and cumulants

First cumulant: mean-value parameter

$$\mu(\theta) = E(Y;\theta) = \int y f_\theta(y)\, dy = K_0'(\theta)$$

Second cumulant: variance function

$$\sigma^2(\theta) = \mathrm{var}(Y;\theta) = K_0''(\theta) = V(\mu)$$
$$V(\mu) = d\mu(\theta)/d\theta$$

Convexity: $K_0'' \geq 0$ implies K is convex on Θ

Third cumulant: skewness

$$E((Y-\mu)^3;\theta) = K_0'''(\theta)$$

Fourth cumulant: kurtosis

$$E((Y-\mu)^4) - 3\sigma^4(\theta)$$ Relevance of cumulants in statistics:

Additivity
Deviations from normality

Examples I–II

Example I: $f_0(y) = e^{-y^2/2}/\sqrt{2\pi}$

$$\begin{aligned} M_0(\theta) &= \int e^{\theta y} f_0(y)\, dy = \exp(\theta^2/2) \\ K_0(\theta) &= \theta^2/2 \\ f_\theta(y) &= e^{y\theta - \theta^2/2} f_0(y) = f_0(y-\theta) \end{aligned}$$

$\Theta = \mathcal{R}; \quad K(\theta) = \theta^2/2; \quad \mu = \theta \quad V(\mu) = 1$

Example II: $f_0(y) = e^{-1}/y!$ for $y = 0, 1, \ldots$

$$\begin{aligned} M_0(\theta) &= \sum e^{\theta y} e^{-1}/y! = \exp(e^\theta - 1) \\ K_0(\theta) &= e^\theta - 1 \\ f_\theta(y) &= e^{y\theta + 1 - e^\theta} e^{-1}/y! \\ &= \exp(-e^\theta)(e^\theta)^y/y! = \text{Poisson}(\mu = e^\theta) \end{aligned}$$

$\Theta = \mathcal{R}; \quad K(\theta) = e^\theta; \quad \mu = e^\theta \quad V(\mu) = \mu$

Examples III–IV

Example III: $f_0(y) = \frac{1}{2}\delta_0(y) + \frac{1}{2}\delta_1(y)$

$$\begin{aligned} M_0(\theta) &= \tfrac{1}{2} + \tfrac{1}{2}e^{\theta} \\ K_0(\theta) &= \log(1 + e^{\theta}) - \log(2) \\ f_\theta(y) &= (\delta_0(y) + e^{\theta}\delta_1(y))/(1 + e^{\theta}) \end{aligned}$$

$\Theta = \mathcal{R}; \quad \mu = e^{\theta}/(1 + e^{\theta}) \quad V(\mu) = \mu(1 - \mu)$

Example IV: $f_0(y) = e^{-y}$ for $y > 0$

$$\begin{aligned} M_0(\theta) &= \int_0^\infty e^{\theta y} e^{-y}\, dy = \int e^{(\theta-1)y}\, dy = 1/(1 - \theta) \\ K_0(\theta) &= -\log(1 - \theta) \qquad (\theta < 1) \\ f_\theta(y) &= e^{-(1-\theta)y}(1 - \theta) = \lambda e^{-\lambda y} \end{aligned}$$

$\Theta = (-\infty, 1), \quad K(\theta) = -\log(1 - \theta),$
$\mu = K'(\theta) = 1/(1 - \theta) = 1/\lambda, \quad V(\mu) = K''(\theta) = \mu^2$

Further examples

Example V: Zero-truncated Poisson:

$$
\begin{aligned}
f_0(y) &= 1/y! \quad \text{for} \quad y = 1, 2, \ldots \\
f_\theta(y) &= e^{\theta y - K(\theta)}/y! \\
e^{K(\theta)} &= \sum_{y=1}^{\infty} e^{\theta y}/y! = \exp(e^\theta) - 1 \\
K(\theta) &= \log(\exp(e^\theta) - 1) \\
K'(\theta) &= \exp(e^\theta)/(\exp(e^\theta) - 1) = \lambda/(1 - e^{-\lambda})
\end{aligned}
$$

Commutativity: truncation followed by exponential weighting versus exponential weighting followed by truncation

Also: Gamma, multinomial, non-central hypergeometric

Circular data

Example VI: $f_0(x) = 1/(2\pi)$ for $|x| = 1$ in $\mathcal{R}^2$.

$$\begin{aligned} y(x) &= \cos(x): \quad \text{canonical statistic} \\ f_\theta(x) &\propto e^{\theta\cos(x)} f_0(x) = e^{\theta\cos(x)}/(2\pi I_0(\theta)) \end{aligned}$$

von Mises-Fisher distribution:

Moment generating function of Y is $I_0(\theta)$
Cumulant generating function is $\log I_0(\theta)$

Similar distributions on the sphere...

More arcane examples

Example VII: $f_0(x)$: uniform on symmetric group S_n

$$\begin{aligned} f_0(x) &= 1/n! \quad (x \in S_n) \\ \text{statistic:} \quad y(x) &= \#x = \text{No of cycles in } x \\ f_\lambda(x) &= \frac{\lambda^{\#x}\Gamma(\lambda)}{\Gamma(n+\lambda)} \\ &= e^{y(x)\theta - K(\theta)} f_0(x) \end{aligned}$$

$$\lambda = e^\theta, \quad K(\theta) = \log\Gamma(e^\theta) + \log\Gamma(n+1) - \log\Gamma(n+e^\theta)$$

Induced distribution on partitions of $[n]$ (Ewens, 1972)

$$f_\lambda(B) = \frac{\lambda^{\#B}\Gamma(\lambda)}{\Gamma(n+\lambda)} \prod_{b\in B} \Gamma(\#b)$$

Cumulant generating function of $\#B$ is $K(\theta)$

$E(\#B) = K'(\theta), \quad \text{var}(\#B) = K''(\theta), \ldots$

Convolution and exponential tilting

Y has density f on $\mathcal{R}$;
$Y_1, \ldots, Y_\nu$ independent and identically distributed copies
Sum $Y = (Y_1 + \cdots + Y_\nu)$ has density $f^{\star\nu}$

$$
\begin{aligned}
f^{\star\nu}(y) &= \int\!\!\int_{\mathbf{y}:\mathbf{1}'\mathbf{y}=y} f(y_1)\cdots f(y_\nu)\, d\mathbf{y} \\
(f_\theta)^{\star\nu}(y) &= \int\!\!\int_{\mathbf{y}:\mathbf{1}'\mathbf{y}=y} f_\theta(y_1)\cdots f_\theta(y_\nu)\, d\mathbf{y} \\
&= \int\!\!\int_{\mathbf{y}:\mathbf{1}'\mathbf{y}=y} e^{\theta y_1 - K(\theta)} f(y_1)\cdots e^{\theta y_\nu - K(\theta)} f(y_\nu)\, d\mathbf{y} \\
&= e^{\theta y - \nu K(\theta)} \int\!\!\int_{\mathbf{y}:\mathbf{1}'\mathbf{y}=y} f(y_1)\cdots f(y_\nu)\, d\mathbf{y} \\
&= e^{\theta y - \nu K(\theta)} f^{\star\nu}(y) = (f^{\star\nu})_\theta(y)
\end{aligned}
$$

Averages

Average has density $\nu f^{\star\nu}(\nu y)$ at y
Exponentially tilted form has density

$$\nu f_\theta^{\star\nu}(\nu y) = e^{\nu(y\theta - K(\theta))} \times \nu\, f^{\star\nu}(\nu y)$$

Can now treat (θ, ν) as a parameter pair

$\bar{Y} \sim \nu f_\theta^{\star\nu}(\nu \cdot)$
$\mu = E(\bar{Y}) = K'(\theta)$,
$\mathrm{var}(\bar{Y}) = K''(\theta)/\nu = V(\mu)/\nu = \sigma^2 V(\mu)$
with $\sigma^2 = 1/\nu$ as dispersion parameter

Maximum likelihood estimation

Model components:

$$E(Y_i) = \mu_i; \quad \eta_i = g(\mu_i) = x_i'\beta$$
$$\sigma^2 = \mathrm{var}(Y_i)/V(\mu_i) = 1/\nu = \text{const}$$

Contribution to the log likelihood from ith component:

$$\begin{aligned}
l_i &= \nu(y_i\theta_i - K(\theta_i)) \\
\frac{\partial l_i}{\partial \theta_i} &= \nu(y_i - K'(\theta_i)) = \nu(y_i - \mu_i) \\
\frac{\partial l_i}{\partial \mu_i} &= \nu(y_i - \mu_i)\frac{d\theta_i}{d\mu_i} = \nu\frac{y_i - \mu_i}{V(\mu_i)} \\
\frac{\partial l_i}{\partial \beta_r} &= \nu x_{ir} g'(\mu_i)\frac{y_i - \mu_i}{V(\mu_i)} \\
-E\left(\frac{\partial^2 l_i}{\partial \beta_r \partial \beta_s}\right) &= \nu x_{ir}x_{is}(g'(\mu_i))^2/V(\mu_i) = \nu x_{ir}x_{is}W_i
\end{aligned}$$

Maximum likelihood estimation

Matrix form of derivatives:

$$\frac{\partial l}{\partial \beta} = \nu X' W Z$$

$$-E\left(\frac{\partial^2 l}{\partial \beta^2}\right) = \nu X' W X \qquad \text{(Fisher information)}$$

$$Z_i = ((y_i - \mu_i)/g(\mu_i)); \qquad W = \text{diag}\{g'(\mu_i)^2/V(\mu_i)\}$$

Newton-Raphson step from $\hat{\beta}^{(0)}$:

$$\hat{\beta}^{(1)} = \hat{\beta}^{(0)} + (X' W X)^{-1} X' W Z$$

Iterate until convergence (4-8 cycles usually suffices)
Value of $\sigma^2 = 1/\nu$ is immaterial

Asymptotic distribution theory

Parameter estimates:

$$E(\hat{\beta}) = \beta + O(n^{-1})$$
$$\mathrm{cov}(\hat{\beta}) = \sigma^2(X'WX)^{-1}(1 + O(n^{-2}))$$
$$\hat{\beta} - \beta \sim N(0, \sigma^2(X'WX)^{-1})$$

under reasonable conditions on X as $n \to \infty$.

Estimation of dispersion parameter:

$$\begin{aligned} R_i &= \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \\ X^2 &= \sum R_i^2 = \sum \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \\ s^2 &= X^2/(n-p) \simeq \sigma^2 \end{aligned}$$

Deviance

Set $\nu = 1$, $\sigma^2 = 1$
Log likelihood at point μ:

$$l(\mu; y) = \sum_i y_i \theta(\mu_i) - K(\theta(\mu_i))$$

$K'(\theta) = \mu$ implies $\theta(\mu) = (K')^{-1}(\mu)$
Log likelihood attained by fitted model

$$l(\hat{\mu}; y) = \sum y_i \theta(\hat{\mu}_i) - K(\theta(\hat{\mu}_i))$$

Largest attainable unconstrained value occurs at $\tilde{\mu} = y$
Deviance at μ = 2*Deficiency = $2l(\tilde{\mu}; y) - 2l(\mu; y)$ is positive

For model comparisons with fixed ν:
(Deviance reduction) $\times \nu$ = Twice log likelihood increase

Deviance

Explicit functional forms for the deviance

Gaussian: $D(\mu; y) = \sum(y_i - \mu_i)^2$

Poisson(μ): $2(y \log(y/\mu) - (y - \mu)), \quad (y \geq 0)$

Binomial(m, π): $2y \log(y/\mu) + 2(m - y) \log((m - y)/(m - \mu))$, $\mu = E(Y) = m\pi$

Exponential/Gamma: $-2 \log(y/\mu) + 2(y - \mu)/\mu, \quad y > 0$