Generalized linear models III
Log-linear and related models

Peter McCullagh

Department of Statistics
University of Chicago

Polokwane, South Africa
November 2013
Outline

Log-linear models

Binomial models

Hypergeometric models

Case-control studies

Over-dispersion and beta-binomial models
Counted data

Counted data: $Y$ is the number of events observed

First possibility: $i$ is an individual/patient/plot/time interval
   $Y_i$ is the number of events observed for $i$
   Number of children of $i$; emergency-room visits by $i$, ...
   Number of distinct plant species in plot $i$...

   $i$ particular bus on Route 217
   $Y_i$ is the number of passengers at a specific point

   $i$ is a particular male fly
   $Y_i$ is number of occasions courting $F$ rather than $U$

   $Y_t = \#\text{bankruptcies in year } t$

In a Markov chain $i = (r, s)$ is a pair of states
   $Y_i = \#\text{transitions } r \to s$

   $i = (r, s)$ is a pair of journals
   $Y_i = \#\text{citations by } r \text{ to } s$
Counted data contd

Each event has a type or mark or label
\( Y_x \) is the number of events of type \( x \)

Each citation has a mark \( x = (r, s) \)
\( Y_x = \# \) marks of type \( x \)

Each toothpaste purchase has a brand
\( Y_x = \# \) units of brand \( x \) purchased

Every traffic accident has a classification
fatal/not; alcohol/not; insurance/not;...
\( Y_x \) is number of accidents of type \( x \)

Same for crimes, web clicks, ...

\( x = (year, state) \): \( Y_x = \# \) hurricanes

PET: \( Y_x \) photon count at angle \( x \)
Poisson models

\( Y_{x,t} \): number of events of type \( x \) observed in \((0, t)\) values recorded for a range of types and intervals

Poisson model: \( Y_{x,t} \sim \text{Po}(\rho_x \times t) \)
- Independent for types \( x \neq x' \)
- Independent for non-overlapping intervals \((0, s), (s, t)\)

stationarity: \( \rho_x \) is the rate for type \( x \)

Log-linear models:
- several classification factors \( x = (r, s, u) \):
  \[
  \log \rho_x = \alpha_r + \beta_s + \gamma_u \quad \text{(w/o) interaction}
  \]
  \[
  E(Y_{x,t}) = e^{\alpha_r} e^{\beta_s} e^{\gamma_u} \times t
  \]
- \( \log \rho_x = \alpha_{rs} + \beta_{st} \) with interaction
Model: $Y_i \sim \text{Po}(\rho_i \times t_i)$ with independent components
\[
\log \mu = X\beta + \log(t)
\]
\[
X'(Y - \hat{\mu}) = 0 \quad \text{(local max condition)}
\]
\[
\hat{\beta} - \hat{\beta}_0 = (X' \hat{W}_0 X)^{-1} X'(Y - \hat{\mu}_0)
\]
\[
\hat{\mu} = \exp(X\hat{\beta}_1 + \log(t)) \quad \text{(one iteration)}
\]
\[
W = \text{diag}\{\mu_1, \ldots, \mu_n\}
\]
\[
\text{cov}(\hat{\beta}) = (X' WX)^{-1}
\]

Implications of the condition $X'Y = X'\hat{\mu}$

Technical point:
- If $n = 1$, and $y = 0$: zero events observed in $(0, t]$
- then $\hat{\mu} = 0$, $\hat{\beta} = -\infty$; technically no mle
- same thing can occur more generally
**Over-dispersed counts**

Practical point:
- Often there are reasons to think that $\text{var}(Y_i) > \mu_i$
- Often but not always
- Can check by computing $\hat{\sigma}^2 = \chi^2 / (n - p)$
- And referring to null distribution $\chi_{n-p}^2 / (n - p)$

Causes of over-dispersion:
- Unrecorded effects, (randomly varying intensity)
- Superposition (double counting);
  - Individuals versus families
  - Individuals versus twins

Remedies:
- Accommodate using over-dispersion parameter
- Negative binomial model

Under-dispersion:
- Much less common, but can occur
- Same event counted in several cells (types not disjoint)
Multinomial models

Poisson to multinomial:

\[ Y_1, \ldots, Y_k \text{ indep Poisson, means } \mu_1, \ldots, \mu_k \]

Conditional distribution given \( Y. = m \):

\[
p(y) = p(y | Y. ) \times p(Y.) \]

\[
e^{-\mu} \cdot \frac{\mu^{y_1}}{y_1!} \cdots \frac{\mu^{y_k}}{y_k!} = \frac{y. !}{y_1! \cdots y_k!} \pi_1^{y_1} \cdots \pi_k^{y_k} \times e^{-\mu} \cdot \frac{\mu^y}{y. !}
\]

\[ \pi_r = \mu_r / \mu. \]

Poisson log lik = multinomial log lik + Poisson log lik for sum

Can fit a multinomial model artificially using a Poisson formulation

Not necessary, but an option...
Binomial models

\[ p(y; m) = \binom{m}{y} \pi^y (1 - \pi)^{m-y} \quad 0 \leq y \leq m \]

\[ \log p = y \log \left( \frac{\pi}{1 - \pi} \right) + m \log(1 - \pi) + \log \left( \binom{m}{y} \right) \]

\[ \theta = \log \left( \frac{\pi}{1 - \pi} \right); \quad \pi = e^\theta / (1 + e^\theta); \]
\[ K(\theta) = -m \log(1 - \pi) = m \log(1 + e^\theta) \]

Linear logistic models: \( \theta \equiv \eta = X\beta \)

Probit models: \( \eta \equiv \Phi^{-1}(\pi) = X\beta \)

C-log log model: \( \eta \equiv \log(-\log(1 - \pi)) = X\beta \)
Parameter interpretation: logistic models

Linear logistic model:

\[
\begin{align*}
    \text{pr}(Y = 1 \mid x) &= \pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \\
    \text{odds}(Y = 1 \mid x) &= \frac{\pi(x)}{1 - \pi(x)} = e^{\beta_0 + \beta_1 x} \\
    \eta(x) &= \log\text{odds}(Y = 1 \mid x) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) \\
    \eta(x) &= \beta_0 + \beta_1 x
\end{align*}
\]

\(\beta\) is the change in log odds per unit change in \(x\)

\(e^\beta\) is the multiplicative change in odds per unit change in \(x\)

\[
\text{odds}(Y = 1 \mid x) = e^\beta \text{odds}(Y = 1 \mid x - 1)
\]

Often easier to talk in terms of odds or risks
2 × 2 hypergeometric models

Design in matched blocks:
- $m_{b0}$ controls, $m_{b1}$ treated individuals in block $b$
- Success probabilities in block $b$: $\pi_{b0}$ and $\pi_{b1}$
  - All responses independent!

Observed counts:
- $Y_0 \sim B(m_0, \pi_{b0})$, $Y_1 \sim B(m_1, \pi_{b1})$

Model: $\log(\text{odds}(\text{success} \mid b, t)) = \alpha_b + \beta_t$

Model: $\text{odds}(\text{success} \mid \text{treat}) = e^\beta \text{odds}(\text{success} \mid \text{cntrl})$

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>S</th>
<th>Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>$m_0 - Y_0$</td>
<td>$Y_0$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>treated</td>
<td>$m_1 - Y_1$</td>
<td>$Y_1$</td>
<td>$m_1$</td>
</tr>
</tbody>
</table>

Conditional distn of $Y_1$ given $Y_0 + Y_1 = s$ is hypergeometric

$$\text{pr}(Y_1 = y \mid m_0, m_1, s) \propto \binom{m_0}{s-y} \binom{m_1}{y} e^{\beta y}$$

Independent of block parameter $\alpha_b$

Exponential family with canonical parameter $\beta$
Case-control studies: retrospective sampling

\( x \): exposure level 0 or 1
\( Y \): healthy or diseased, 0 or 1; (rare diseases)

\[
\text{pr}(Y = 1 \mid x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}
\]

Response-dependent sampling fractions:

\[
\begin{align*}
\text{pr}(Z_i = 1 \mid x_i, Y_i = 0) & = \pi_0 & \text{(for controls)} \\
\text{pr}(Z_i = 1 \mid x_i, Y_i = 1) & = \pi_1 & \text{(for cases)}
\end{align*}
\]

By Bayes’s theorem

\[
\text{pr}(Y = 1 \mid Z = 1, x) = \frac{\text{pr}(Z = 1 \mid Y = 1, x) \ \text{pr}(Y = 1 \mid x)}{\cdots + \text{pr}(Z = 1 \mid Y = 0, x) \ \text{pr}(Y = 0 \mid x)} = \pi_1 e^{\alpha + \beta x} / (\pi_0 + \pi_1 e^{\alpha + \beta x})
\]

\[
\text{logit } \pi(Y = 1 \mid Z = 1, x) = \alpha + \log(\pi_1 / \pi_0) + \beta x
\]

same \( \beta \), different intercept
beta-binomial models

Example: antibody counts for 18 patients

<table>
<thead>
<tr>
<th>pid</th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>c6</th>
<th>c7</th>
<th>c8</th>
<th>s1</th>
<th>s2</th>
<th>s3</th>
<th>s4</th>
<th>s5</th>
<th>s6</th>
<th>s7</th>
<th>s8</th>
<th>s9</th>
<th>s10</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHA</td>
<td>13</td>
<td>12</td>
<td>39</td>
<td>5</td>
<td>9</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>13</td>
<td>18</td>
<td>8</td>
<td>24</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Tot</td>
<td>20</td>
<td>18</td>
<td>43</td>
<td>6</td>
<td>15</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>16</td>
<td>24</td>
<td>14</td>
<td>30</td>
<td>12</td>
<td>11</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Dispersion: $X^2 = 23.6$ for controls and 14.3 for SLE patients.
Lexis dispersion index: $37.9/16 = 2.4$

beta-binomial distribution

$$
\binom{m}{y} \frac{\alpha^y \beta^{m-y}}{(\alpha + \beta)^m} \quad (0 \leq y \leq m)
$$

where $\alpha^\up{n} = \alpha(\alpha + 1) \cdots (\alpha + n - 1)$ is the ascending factorial.

$E(Y) = m\alpha/(\alpha + \beta) = m\pi$

$\text{var}(Y) = m\pi(1 - \pi)(\alpha + \beta + m)/(\alpha + \beta + 1)$

beta-binomial = binomial = bernoulli if $m = 1$
Beta-binomial applied to antibody counts

Parameter estimates (max likelihood)

<table>
<thead>
<tr>
<th></th>
<th>$H_0$</th>
<th>$H_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_c$</td>
<td>8.437</td>
<td>5.057</td>
</tr>
<tr>
<td>$\beta_c$</td>
<td>3.392</td>
<td>2.588</td>
</tr>
<tr>
<td>$\alpha_s$</td>
<td>8.437</td>
<td>26.705</td>
</tr>
<tr>
<td>$\beta_s$</td>
<td>3.392</td>
<td>11.382</td>
</tr>
<tr>
<td>$2 \times llk$</td>
<td>$-298.31$</td>
<td>$-297.38$</td>
</tr>
</tbody>
</table>

Binomial log likelihoods: $-305.38$ for $H_0$ and $-305.36$ for $H_a$

Conclusion:
Significant over-dispersion (LR = 7.1 on 1 df)
but no difference between patients types S and C

$\hat{\pi} = \hat{\alpha}/(\hat{\alpha} + \hat{\beta}) = 0.71$; se = ??