Estimating Overdispersion in Sparse Multinomial Data

David Fletcher
Department of Mathematics and Statistics
University of Otago
Dunedin
New Zealand

July 23, 2013
Overview

- Motivating Example
- Lack-of-Fit in Sparse Multinomial Data
- Overdispersion and Quasi-likelihood
- Use of Results for Count Data
- Estimation of Overdispersion Parameter
- Asymptotics and Simulations
Motivating Example

Mark-Recapture Data

- Banding of swifts (*Apus apus*) in Nîmes
- Eight-year study of 93 individuals
Motivating Example

Mark-Recapture Data

- Banding of swifts (Apus apus) in Nîmes
- Eight-year study of 93 individuals
- Capture histories (1 if captured, 0 otherwise)

\[
\begin{align*}
11111110 \\
11101100 \\
\ldots \\
01110100 \\
01110000 \\
\ldots \\
00000001 \\
00000001
\end{align*}
\]
Motivating Example

Mark-Recapture Data

- 75 individuals provide information on (apparent) survival
- 254 possible capture histories; 87% have a frequency of zero
- Sparse multinomial data

<table>
<thead>
<tr>
<th>Release Cohort</th>
<th>Capture History</th>
<th>Frequency (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100000000</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>100000001</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>01000000</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>01000001</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>7</td>
<td>00000010</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>00000011</td>
<td>8</td>
</tr>
</tbody>
</table>
Model

- For each release cohort
  
  \[(Y_1, \ldots, Y_n) \sim \text{Multinomial}(N, \pi)\]

- Product-multinomial model

- \(n\) and \(N\) vary by cohort; \(\pi\) usually does not

- \(\pi\) modelled using \(p\) parameters for survival and recapture

- Sparseness: many \(\pi\) close to zero
Lack-of-Fit

- Lack-of-fit of product-multinomial model?
- Random effects vs Quasi-likelihood

Quasi-likelihood

\[
\text{var}(Y_i) = \phi V_i
\]

\(V_i = \text{variance predicted by model}\)
\(\phi = \text{overdispersion parameter}\)

Estimate of \(\phi \Rightarrow \text{CIs, model selection/averaging}\)
Multinomial vs Poisson

- Distribution of $\hat{\phi}$ given estimates of other parameters
- Consider $\hat{\phi}$ for corresponding Poisson model (with “intercept”)
- (c.f. use of log-linear models for contingency tables)

Quasi-likelihood for Count Data:

$$E(Y_i) = \mu_i \quad \text{var}(Y_i) = \phi \mu_i$$

Wedderburn (1974):

$$\hat{\phi}_1 = \frac{X^2}{n - p} \quad X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Highly variable when data sparse ($\mu_i$ small)
Swift Example

Data: $\hat{\phi}_1 = 1.69$

Simulations:

- Each cohort:
  
  $$(Y_1, \ldots, Y_n) \sim \text{Dirichlet-multinomial}(N = 10, \hat{\pi}, \phi = 2)$$

- Bias = -0.78, SE = 1.78, RMSE = 1.96 (10^3 simulations)
Farrington (1995) suggested the following (for testing $\phi = 1$):

$$
\hat{\phi}_2 = \hat{\phi}_1 - \sum_i \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \frac{1}{n - p}
$$

Asymptotics: less variable than $\hat{\phi}_1$ when $\mu_i$ small
Overdispersion in Count Data

Estimating equation:

\[ \sum_i [a_i (y_i - \hat{\mu}_i) + b_i \{(y_i - \hat{\mu}_i)^2 - \phi \hat{\mu}_i\}] = 0 \]

- \( \hat{\mu}_i \) = quasi-likelihood estimate of \( \mu_i \) (identical to MLE)
- Quadratic estimating equations (Crowder 1987)

Choose \( a_i = a/\hat{\mu}_i \) and \( b_i = 1/\hat{\mu}_i \) with:

<table>
<thead>
<tr>
<th>Estimator</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wedderburn</td>
<td>0</td>
</tr>
<tr>
<td>Farrington</td>
<td>-1</td>
</tr>
<tr>
<td>New</td>
<td>-( \phi )</td>
</tr>
</tbody>
</table>
Theorem: If third cumulant of $Y = \alpha \mu_i$ ($\alpha \geq \phi^2$) then new estimator has the lowest variance.

Condition on third cumulant satisfied by:

- Poisson stopped-sum distributions (negative binomial, Neyman Type A, Polya-Aeppli and Hermite)
- Poisson-lognormal
- Poisson-inverse-Gaussian

Bias also affected by choice of $\alpha$. 

Estimators

\[ \hat{\phi}_1 = \frac{X^2}{n - p} \]

\[ \hat{\phi}_2 = \hat{\phi}_1 - \frac{n\bar{s}}{n - p} \]

\[ \hat{\phi}_3 = \frac{\hat{\phi}_1}{1 + \bar{s}} \]

\[ X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \]

\[ \bar{s} = \frac{1}{n} \sum_i \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \]
Swift Example

Focus on $\hat{\phi}_1$ vs $\hat{\phi}_3$

Data: $\hat{\phi}_1 = 1.69$, $\hat{\phi}_3 = 1.06$

Simulations:

<table>
<thead>
<tr>
<th></th>
<th>$N = 10$ per cohort</th>
<th>$N = 100$ per cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SE</td>
</tr>
<tr>
<td>$\hat{\phi}_1$</td>
<td>-0.78</td>
<td>1.78</td>
</tr>
<tr>
<td>$\hat{\phi}_3$</td>
<td>-0.98</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Parametric bootstrap bias-adjustment

\((Y_1^*, \ldots, Y_n^*) \sim \text{Dirichlet-multinomial}(N, \hat{\pi}, \hat{\phi})\)

Estimate of relative bias = \(\bar{\phi}^* / \bar{\phi}\)

\(\bar{\phi}^* = \text{mean of bootstrap estimates}\)

Bias-adjusted estimator: \(\hat{\phi}_B = \hat{\phi}^2 / \bar{\phi}^*\)

NB Modify slightly if \(\hat{\phi} < 1\) or \(\hat{\phi} > N\)
### Parametric bootstrap bias-adjustment

$10^3$ bootstrap samples

<table>
<thead>
<tr>
<th></th>
<th>$N = 10$ per cohort</th>
<th></th>
<th></th>
<th>$N = 100$ per cohort</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SE</td>
<td>RMSE</td>
<td>Bias</td>
<td>SE</td>
<td>RMSE</td>
</tr>
<tr>
<td>$\hat{\phi}_1$</td>
<td>-0.78</td>
<td>1.78</td>
<td>1.95</td>
<td>0.00</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>$\hat{\phi}_3$</td>
<td>-0.98</td>
<td>0.42</td>
<td>1.06</td>
<td>-0.06</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>$\hat{\phi}_{3B}$</td>
<td>0.08</td>
<td>0.72</td>
<td>0.72</td>
<td>0.12</td>
<td>0.74</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Bias vs RMSE when estimating a dispersion parameter
Further Work

- Phylogenetics (n huge; highly sparse)
- Bootstrap adjustment without re-fitting model?
- Confidence interval for $\phi$
References

- Crowder (1987) Biometrika 74, 591-7
- Farrington (1996) JRSSB 58, 349-60
- Fletcher (submitted) Biometrics