A New Method for Estimating Overdispersion in Count Data

David Fletcher

May 5, 2011
Overview

- Motivation
- Example
- Overdispersion in count data
- Current methods for estimating overdispersion
- Quadratic estimating equations
- New method
- Simulations
- Further work
Motivation

- Mark-recapture data (multinomial likelihood)
- Lack-of-fit diagnostics: how can we improve the model?
- Overdispersion: variation greater than predicted
- Count data: estimating overdispersion tricky when mean is low
- Bootstrap-based method?
- Farrington’s modification to Pearson’s lack-of-fit statistic
- Estimating overdispersion vs testing lack-of-fit
Example: sealion bycatch

<table>
<thead>
<tr>
<th></th>
<th>Sealions killed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Spring</td>
<td>20</td>
</tr>
<tr>
<td>Summer</td>
<td>87</td>
</tr>
<tr>
<td>Autumn</td>
<td>125</td>
</tr>
</tbody>
</table>

- Bycatch of New Zealand sealions in squid fishery (1990 – 1996)
- Sample unit: all tows by one vessel in six hours
- Offset can allow for number of tows per sample unit (1 – 40)
- Low mean count
Overdispersion in count data

- Poisson model

\[ Y_i \sim \text{Poisson}(\mu_i) \]
\[ E(Y_i) = V(Y_i) = \mu_i \]
\[ g(\mu_i) = X\beta \]

- Variance often greater than the mean (overdispersion)
- SEs too small and overly-complex models selected
- Use estimate of $\phi$ to adjust SEs and model selection
Overdispersion in count data

- Quasi-Poisson approach (special case of quasi-likelihood):

\[ E(Y_i) = \mu_i \quad V(Y_i) = \phi \mu_i \]

where \( \phi (\ > 1) \) is the dispersion parameter

- Leads to same \( \hat{\beta} \) as Poisson model

- Negative binomial model:

\[ E(Y_i) = \mu_i \quad V(Y_i) = \mu_i + \frac{\mu_i^2}{k_i} \]

\[ k_i = \mu_i / (\phi - 1) \Rightarrow V(Y_i) = \phi \mu_i \]
Current methods for estimating $\phi$

- **Quasi-Poisson approach:**

$E(Y_i) = \mu_i \quad V(Y_i) = \phi \mu_i$

- **Wedderburn 1974 proposed the current standard:**

$$\hat{\phi}X = \frac{X^2}{n - p}$$

where

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

- **Prone to problems when $\mu_i$ small:** highly variable
Current methods for estimating $\phi$

- Farrington (1995) proposed:

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

- Asymptotic theory suggests $\hat{\phi}_F$ better than $\hat{\phi}_X$, especially when $\mu_i$ small: less variable and lower correlation with $\hat{\beta}$
Crowder (1987) proposed the general use of quadratic estimating equations. For overdispersed count data these are:

\[ g_r(\theta) = \sum_i \left[ a_i^{(r)}(y_i - \mu_i) + b_i^{(r)} \left\{ (y_i - \mu_i)^2 - \phi \mu_i \right\} \right] = 0 \]

for \( r = 1, \ldots, p + 1 \), where \( \theta = (\beta, \phi)^T \) and the \( a_i^{(r)}, b_i^{(r)} \) are functions of \( \theta \).

- Special cases: least squares and quasi-likelihood (\( b_i^{(r)} = 0 \))
- Partly motivated by issues with quasi-likelihood, including desire to use information on \( \beta \) contained in the variance
Quadratic estimating equations

- For overdispersed count data these are:

\[ g_r (\theta) = \sum_i \left[ a_i^{(r)} (y_i - \mu_i) + b_i^{(r)} \left\{ (y_i - \mu_i)^2 - \phi \mu_i \right\} \right] = 0 \]

for \( r = 1, \ldots, p + 1 \), where \( \theta = (\beta, \phi)^T \) and the \( a_i^{(r)} \), \( b_i^{(r)} \) are functions of \( \theta \)

- Optimal choices available for \( a_i^{(r)} \) and \( b_i^{(r)} \) IF we specify third and fourth cumulants

- Optimal in the sense that

\[ V_0 \left( \hat{\theta}^* \right) - V_0 \left( \hat{\theta} \right) \]

is positive-semidefinite, where \( \hat{\theta}^* \) is obtained using any other choice of \( a_i^{(r)} \) and \( b_i^{(r)} \)
Optimal $a_i^{(r)}$ and $b_i^{(r)}$

For $r = 1, \ldots, p$

$$a_i^{(r)} = \left( \frac{\phi^{1/2} \mu_i^{-1/2} \gamma_1 - \gamma_2 - 2}{\phi \mu_i (\gamma_2 - \gamma_1^2 + 2)} \right) \frac{\partial \mu_i}{\partial \beta_r}$$

$$b_i^{(r)} = \left( \frac{\gamma_1 - \phi^{1/2} \mu_i^{-1/2}}{\phi^{3/2} \mu_i^{3/2} (\gamma_2 - \gamma_1^2 + 2)} \right) \frac{\partial \mu_i}{\partial \beta_r}$$

$$a_i^{(p+1)} = \frac{\gamma_1 \phi^{-3/2} \mu_i^{-1/2}}{\gamma_2 - \gamma_1^2 + 2}$$

$$b_i^{(p+1)} = -\frac{\phi^{-2} \mu_i^{-1}}{\gamma_2 - \gamma_1^2 + 2}$$

$\gamma_1 i = \kappa_3 i / \kappa_2^{3/2}$, $\gamma_2 i = \kappa_4 i / \kappa_2^2$ and $\kappa_{ji}$ is the $j^{th}$ cumulant of $Y_i$

$$\left( \kappa_1 i = \mu_i \quad \kappa_2 i = \mu_2 i \quad \kappa_3 i = \mu_3 i \quad \kappa_4 i = \mu_4 i - 3 \mu_2^2 i \right)$$
New method of estimating $\phi$

If there were a Quasi-Poisson "distribution" in the exponential family (i.e. with $V(Y_i) = \phi \mu_i$) it would satisfy

$$\kappa_{3i} = \phi^2 \kappa_{3i}^* \quad \text{and} \quad \kappa_{4i} = \phi^3 \kappa_{4i}^*$$

where $\kappa_{3i}^* = \mu_i$ and $\kappa_{4i}^* = \mu_i$ are the third and fourth cumulants of the Poisson($\mu_i$) distribution.

If we assume these third and fourth cumulants, the optimal $a_i^{(r)}$ and $b_i^{(r)}$ lead to the same $\hat{\beta}$ as for the Quasi-Poisson (and Poisson) approach and also give

$$\tilde{\phi} = \frac{\chi^2}{\sum_i \frac{Y_i}{\hat{\mu}_i}}$$
Comparison of methods

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

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

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

\[
\hat{\phi} = \frac{X^2}{\sum_i \frac{y_i}{\hat{\mu}_i}}
\]
Example: sealion bycatch

<table>
<thead>
<tr>
<th></th>
<th>Sealions killed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0   1   2   3</td>
</tr>
<tr>
<td>Spring</td>
<td>20  1   0   0</td>
</tr>
<tr>
<td>Summer</td>
<td>87  3   0   0</td>
</tr>
<tr>
<td>Autumn</td>
<td>125 27  4   2</td>
</tr>
</tbody>
</table>

- Poisson model with *season* as a predictor
- Log(number of tows for that sample unit) as an offset
- No evidence of zero-inflation relative to a Poisson model

\[
\hat{\phi}_X = 2.17 \quad \hat{\phi}_F = 1.19 \quad \tilde{\phi} = 1.09
\]
Simulations

- Based on the bycatch example (same sample sizes and number of tows per sample unit)
- $\beta$ set equal to $\hat{\beta}$
- Specified true value for $\phi$ between 1 and 3

- Generated data using one of three models:
  1. Negative binomial with $V(Y_i) = \phi \mu_i$, i.e. $k_i = \mu_i / (\phi - 1)$
  2. Neyman Type A model: $Y_i = \sum_{j=1}^{N_i} Z_{ij}$, where $Z_{ij} \sim Poisson(\phi - 1)$ and $N_i \sim Poisson\left(\frac{\mu_i}{\phi - 1}\right)$
  3. For $\phi = 1$, $Y_i \sim Poisson(\mu_i)$
Simulation results: Negative Bernoulli
Simulation results: Negative binomial

![Graph showing simulation results for Negative binomial distribution with various parameters and their estimates.]
Simulation results: Negative binomial

![Graph showing simulation results for different values of φ.

Key:
- $\hat{\phi}_X$
- $\hat{\phi}_F$
- $\tilde{\phi}$
- Negbin MLE

RMSE vs φ for different values of φ.

Values:
- RMSE
- φ

Graph ranges:
- RMSE from 0.5 to 2.0
- φ from 1.0 to 3.0

Legend:
- Negbin MLE
- $\hat{\phi}_X$
- $\hat{\phi}_F$
- $\tilde{\phi}$
Simulation results: Neyman Type A

![Graph showing simulation results for Neyman Type A with various lines representing different approximations and their biases.]

- Negbin MLE
- \( \hat{\phi}_X \)
- \( \hat{\phi}_F \)
- \( \tilde{\phi} \)
Simulation results: Neyman Type A
Simulation results: Neyman Type A

![Graph showing simulation results for Neyman Type A with various markers and lines representing different estimators.](image-url)
Simulation results
Further Work

- Robustness to assumptions about third and fourth cumluants?
- Asymptotic theory
- Binomial data
- Mark-recapture (multinomial) data
- Quadratic variance-mean relationship?
- Residual diagnostics?
- Bayesian model-checking
Further Work: Robustness to assumptions

\[ \frac{\kappa_3}{\mu} \]

Graph showing the relationship between \( \phi \) and \( \frac{\kappa_3}{\mu} \) with different assumptions:
- Red line: Negative binomial
- Blue line: Neyman Type A
- Black line: Assumed
Further Work: Robustness to assumptions

$k_4/\mu$

- Red: Negative binomial
- Blue: Neyman Type A
- Black: Assumed
Further Work

- Robustness to assumptions about third and fourth cumulants?
- Asymptotic theory
- Binomial data
- Mark-recapture (multinomial) data
- Quadratic variance-mean relationship?
- Residual diagnostics?
- Bayesian model-checking
Further Work: Residual diagnostics
Further Work

- Robustness to assumptions about third and fourth cumluants?
- Asymptotic theory
- Binomial data
- Mark-recapture (multinomial) data
- Quadratic variance-mean relationship?
- Residual diagnostics?
- Bayesian model-checking