Size Estimation - Statistical Models for Underreporting

Gerhard Neubauer, Gordana Djuraš & Herwig Friedl

JOANNEUM RESEARCH and Technical University, Graz
1 Introduction
Underreporting

Any sample of count data may be incomplete

- criminology: crimes with an aspect of shame (sexuality, domestic violence) or theft of low values goods
- public health: infectious (HIV) or chronic (diabetes) disease data
- production: error counts in a production process
- traffic accidents with minor damage

Estimation of total number of cases
## Binomial Model

Event reported

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td></td>
<td>$1-R$</td>
</tr>
</tbody>
</table>

$R \sim Bernoulli(\pi)$
## Binomial Model

**Event reported**

<table>
<thead>
<tr>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>$1 - R$</td>
</tr>
</tbody>
</table>

$R \sim Bernoulli(\pi)$

iid sample of size $\lambda \Rightarrow Y = \sum_i R_i \sim Binomial(\lambda, \pi)$

$E(Y) = \mu = \lambda \pi, \quad \text{var}(Y) = \sigma^2 = \lambda \pi (1 - \pi)$
**Binomial Model**

Event reported

<table>
<thead>
<tr>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R )</td>
<td>( 1 - R )</td>
</tr>
</tbody>
</table>

\( R \sim Bernoulli(\pi) \)

iid sample of size \( \lambda \Rightarrow Y = \sum_i R_i \sim Binomial(\lambda, \pi) \)

\[
E(Y) = \mu = \lambda \pi, \quad \text{var}(Y) = \sigma^2 = \lambda \pi (1 - \pi)
\]

- \( Y \) the number of reported events
- \( \pi \) the reporting probability
- \( \lambda \) the total number of events - size parameter
- \( U = \lambda - Y \) the number of unreported events
Binomial Model

Event reported

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>1 − $R$</td>
<td></td>
</tr>
</tbody>
</table>

$R \sim \text{Bernoulli}(\pi)$

iid sample of size $\lambda \Rightarrow Y = \sum_i R_i \sim \text{Binomial}(\lambda, \pi)$

$E(Y) = \mu = \lambda \pi, \quad \text{var}(Y) = \sigma^2 = \lambda \pi (1 - \pi)$

Both $\lambda$ and $\pi$ have to be estimated

No longer member of Exponential Family
Estimation

For $T$ iid samples $Y_t$ 

Method of Moments

For the binomial we have $\text{var}(Y) = \mu - \mu^2 / \lambda \leq \mu$

Limitation to data with $s^2 < \bar{y}$

For $s^2 > \bar{y}$

1. Regression approach using ML

   $Y_t \overset{\text{ind}}{\sim} \text{Binomial}(\lambda_t, \pi)$ \textbf{with} $\lambda_t = f(x_t, \beta)$

   Neubauer & Friedl (2006)

2. Mixed model approaches
2 Alternative Models
Beta-Binomial

Random reporting probability $P$

$$Y_t|P \sim \text{Binomial}(\lambda, p)$$

$$P \sim \text{Beta}(\gamma, \delta)$$

$$Y_t \sim \text{Beta-Binomial}(\lambda, \gamma, \delta)$$

Mean-variance relation

$$\text{var}(Y_t) = \left( \mu - \frac{\mu^2}{\lambda} \right) \phi$$

$$\phi = \frac{\lambda + \gamma + \delta}{1 + \gamma + \delta} \geq 1$$
Poisson

Random total number of cases $L$

$$Y_t|L \sim \text{Binomial}(l, \pi)$$

$$L \sim \text{Poisson}(\lambda)$$

$$Y_t \sim \text{Poisson}(\lambda \pi)$$

Parameters not identified
Negative Binomial

Additional randomness in $E(L)$

\[
L | K \sim \text{Poisson}(k \lambda)
\]

\[
Y_t | K \sim \text{Poisson}(k \lambda \pi)
\]

\[
K \sim \text{Gamma}(\omega, \omega)
\]

\[
Y_t \sim \text{Negative Binomial}(\omega, 1 - \pi)
\]

$\omega$ the number of unreported cases  
$\pi$ the reporting probability

Mean-variance relation

\[
\text{var}(Y_t) = \mu + \frac{\mu^2}{\omega} \geq \mu
\]
Beta-Poisson

Consider both binomial parameters as random

\[ Y|L, P \sim \text{Binomial}(L, P) \]
\[ L \sim \text{Poisson}(\lambda) \]
\[ P \sim \text{Beta}(\gamma, \delta) \]
\[ Y \sim \text{Beta-Poisson}(\lambda, \gamma, \delta) \]

\[ E(Y) = \lambda \pi = \mu \quad \text{where} \quad \pi = \frac{\gamma}{\gamma + \delta} \]

\[ \text{var}(Y) = \mu \phi \quad \text{with} \quad \phi = 1 + \frac{\lambda (1 - \pi)}{1 + \gamma + \delta} \geq 1 \]
Generalized Poisson Distribution

Consul (1989)

Moments

\[
\begin{align*}
E(Y) &= \frac{\theta}{(1 - \tau)} \\
\text{var}(Y) &= \frac{\theta}{(1 - \tau)^3}
\end{align*}
\]

- \( \tau = 0 \): \( E(Y) = \text{var}(Y) \) \( \Rightarrow \) Poisson \( (\theta) \)
- \( 0 < \tau < 1 \): \( E(Y) < \text{var}(Y) \) \( \Rightarrow \) Neg. Binomial
- \( \tau < 0 \): \( E(Y) > \text{var}(Y) \) \( \Rightarrow \) Binomial
Conditional Poisson Models

Motivation:

\[ \pi \to 1 \text{ leads to } Y \to \lambda \text{ in the binomial approach} \]

Assume \( Y \mid L \sim Poisson(L) \)

Choose \( p(L) \) such that

\[ E(Y) = \lambda \pi \quad \text{and} \quad \text{var}(Y) = \lambda \pi \phi \]

For example:

\[ L \sim Binomial(\lambda, \pi) \quad 1 < \phi = 2 - \pi < 2 \]

\[ L \sim Negative\ Binomial(\lambda, \pi) \quad 2 < \phi = \frac{2 - \pi}{1 - \pi} < \infty \]
Regression Modelling

For all models - except the GP - we use

$$\lambda_{t,\beta} = \exp(x_t'\beta) \quad \text{and} \quad \pi_\alpha = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$$

$x_t$, $d$-vector of known regressors
$\beta$, $d$-vector of unknown parameters

For the GP model we use

$$\theta_{t,\beta} = \exp(x_t'\beta) \quad \text{and} \quad \tau_\alpha = 1 - \exp(-\alpha)$$

Test $\alpha = 0 \Rightarrow$ Identify Poisson misdispersion
3 Implementation
**R package sizEst**

Full ML estimation of all models: done
Conditional Poisson in work

Testing competing models: in development

Testing parameters within models: done

Model diagnostics: done

Main functions:
arrayEst, sizEst

Implemented methods:
sizEst: plot, predict, residuals, summary
summary.arrayEst
4 Real Data Application
### Stroke Data

**Hypothesis:** Slight strokes are not seen in hospitals

**Data:** Hospital discharges

**Output from function `arrayEst()`**

<table>
<thead>
<tr>
<th></th>
<th>iterations</th>
<th>loglik</th>
<th>chi.sq</th>
<th>gradient</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP</td>
<td>0.5</td>
<td>-405.256</td>
<td>94.602</td>
<td>0.000</td>
<td>0.4594</td>
</tr>
<tr>
<td>NegBin</td>
<td>0.5</td>
<td>-405.195</td>
<td>94.411</td>
<td>0.000</td>
<td>0.4604</td>
</tr>
<tr>
<td>BetaBin</td>
<td>0.5</td>
<td>-402.135</td>
<td>80.779</td>
<td>0.000</td>
<td>0.8230</td>
</tr>
<tr>
<td>BetaPois</td>
<td>0.5</td>
<td>-408.170</td>
<td>80.811</td>
<td>0.000</td>
<td>0.9097</td>
</tr>
</tbody>
</table>
## Stroke Data

Output from function `sizEst()`

**Distribution:** BetaPois  
**Formula:** \( y \sim \beta_0 + T \cdot \cos 1 + T \cdot \sin 1 - 1 \)

| Parameter | Estimate | Std. Error | t value | Pr(>|t|) |
|-----------|----------|------------|---------|----------|
| alpha1    | 2.309    | 0.289      | 8.002   | 0        |
| beta01    | 5.071    | 0.025      | 204.435 | 0        |
| T.cos1    | 0.060    | 0.016      | 3.673   | 0        |
| T.sin1    | -0.083   | 0.016      | -5.249  | 0        |
| Theta     | 11.231   | ---        | ---     | -        |

Performance measures:

<table>
<thead>
<tr>
<th>Measures</th>
<th>values</th>
</tr>
</thead>
<tbody>
<tr>
<td>loglik</td>
<td>-408.170</td>
</tr>
<tr>
<td>chi.sq</td>
<td>80.811</td>
</tr>
<tr>
<td>df.residual</td>
<td>92.000</td>
</tr>
<tr>
<td>aic</td>
<td>824.341</td>
</tr>
<tr>
<td>bic</td>
<td>834.599</td>
</tr>
</tbody>
</table>

Reporting Probabilities:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>lower estimate</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha1</td>
<td>0.8622</td>
<td>0.9571</td>
</tr>
</tbody>
</table>
Stroke Data

Mean function from Quasi-Poisson-Model (blue) and BetaPois model (red)

Lambda function from BetaPois model (red)

Pearson residuals

Histogram of residuals and N(0,1) density

$\hat{\pi} = 0.91$
Crime Register Data

**SIMO**: Austrian online crime register

- Time range: 2004 - 2007
- weekly counts
- 132 regions
- different crime categories

In most cases Poisson overdispersion
GP model estimates

**Shop Lifting**

\[ \hat{\pi} = 0.69 \]

**Bicycle Theft**

\[ \hat{\pi} = 0.67 \]
Beta-Poisson model estimates

\[ \hat{\pi} = 0.65 \]

\[ \hat{\pi} = 0.52 \]
Summary

- Great variety of models
- MLE based implementation in R
- Good performance for simulated data
- Reasonable estimates for real data
Summary

- Great variety of models
- MLE based implementation in R
- Good performance for simulated data
- Reasonable estimates for real data

Future work

- Implement Conditional Poisson models
- Non-nested Testing for more than two models