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

Yes	No
R	$1 - R$

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

Binomial Model

Event reported

Yes	No
R	$1 - R$

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

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

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

Binomial Model

Event reported

Yes	No
R	$1-R$

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

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

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

Y	the number of reported events
π	the reporting probability
λ	the total number of events - **size** parameter
$U = \lambda - Y$	the number of unreported events

Binomial Model

Event reported

Yes	No
R	$1 - R$

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

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

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

Both λ and π 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 $\mathrm{var}(Y) = \mu - \mu^2/\lambda \leq \mu$
Limitation to data with $s^2 < \bar{y}$

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

1. Regression approach using ML
   $Y_t \stackrel{ind}{\sim} Binomial(\lambda_t, \pi)$ with $\lambda_t = f(x_t, \beta)$

   Neubauer & Friedl (2006)

2. Mixed model approaches

2 Alternative Models

Beta-Binomial

Random reporting probability P

$$\begin{aligned} Y_t|P &\sim \text{Binomial}(\lambda, p) \\ P &\sim \text{Beta}(\gamma, \delta) \\ Y_t &\sim \text{Beta-Binomial}(\lambda, \gamma, \delta) \end{aligned}$$

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

$$\begin{aligned} Y_t|L &\sim \text{Binomial}(l, \pi) \\ L &\sim \text{Poisson}(\lambda) \\ Y_t &\sim \text{Poisson}(\lambda\pi) \end{aligned}$$

Parameters not identified

Negative Binomial

Additional randomness in $\mathrm{E}(L)$

$$\begin{aligned} L|K &\sim \text{Poisson}(k\lambda) \\ Y_t|K &\sim \text{Poisson}(k\lambda\pi) \\ K &\sim \text{Gamma}(\omega, \omega) \\ Y_t &\sim \text{Negative Binomial}(\omega, 1-\pi) \end{aligned}$$

ω the number of unreported cases
π the reporting probability

Mean-variance relation

$$\mathrm{var}(Y_t) = \mu + \frac{\mu^2}{\omega} \geq \mu$$

Beta-Poisson

Consider both binomial parameters as random

$$
\begin{aligned}
Y|L,P &\sim \text{Binomial}(L,P) \\
L &\sim \text{Poisson}(\lambda) \\
P &\sim \text{Beta}(\gamma,\delta) \\
Y &\sim \text{Beta-Poisson}(\lambda,\gamma,\delta)
\end{aligned}
$$

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

$$
\text{var}(Y) = \mu\phi \quad \text{with} \quad \phi = 1 + \frac{\lambda(1-\pi)}{1+\gamma+\delta} \geq 1
$$

Generalized Poisson Distribution

Consul(1989)

Moments

$$\mathrm{E}(Y) = \frac{\theta}{(1-\tau)} \qquad \mathrm{var}(Y) = \frac{\theta}{(1-\tau)^3}$$

- $\tau = 0$: $\mathrm{E}(Y) = \mathrm{var}(Y)$ $\Rightarrow$ Poisson (θ)
- $0 < \tau < 1$: $\mathrm{E}(Y) < \mathrm{var}(Y)$ $\Rightarrow$ Neg. Binomial
- $\tau < 0$: $\mathrm{E}(Y) > \mathrm{var}(Y)$ $\Rightarrow$ Binomial

Conditional Poisson Models

Motivation:

$\pi \to 1$ leads to $Y \to \lambda$ in the binomial approach

Assume $Y|L \sim Poisson(L)$

Choose $p(L)$ such that

$$\mathrm{E}(Y) = \lambda\pi \qquad \text{and} \qquad \mathrm{var}(Y) = \lambda\pi\phi$$

For example:

$$L \sim Binomial(\lambda, \pi) \qquad 1 < \phi = 2 - \pi < 2$$

$$L \sim Negative\ Binomial(\lambda, \pi) \qquad 2 < \phi = \frac{2-\pi}{1-\pi} < \infty$$

Regression Modelling

For all models - except the GP - we use

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

x_t, d-vector of known regressors
β, d-vector of unknown parameters

For the GP model we use

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

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()`

```
                   iterations    loglik chi.sq gradient      p1
GP        0.5               3 -405.256 94.602    0.000 0.4594
NegBin    0.5               3 -405.195 94.411    0.000 0.4604
BetaBin   0.5              19 -402.135 80.779    0.000 0.8230
BetaPois  0.5              30 -408.170 80.811    0.000 0.9097
```

Stroke Data

Output from function `sizEst()`

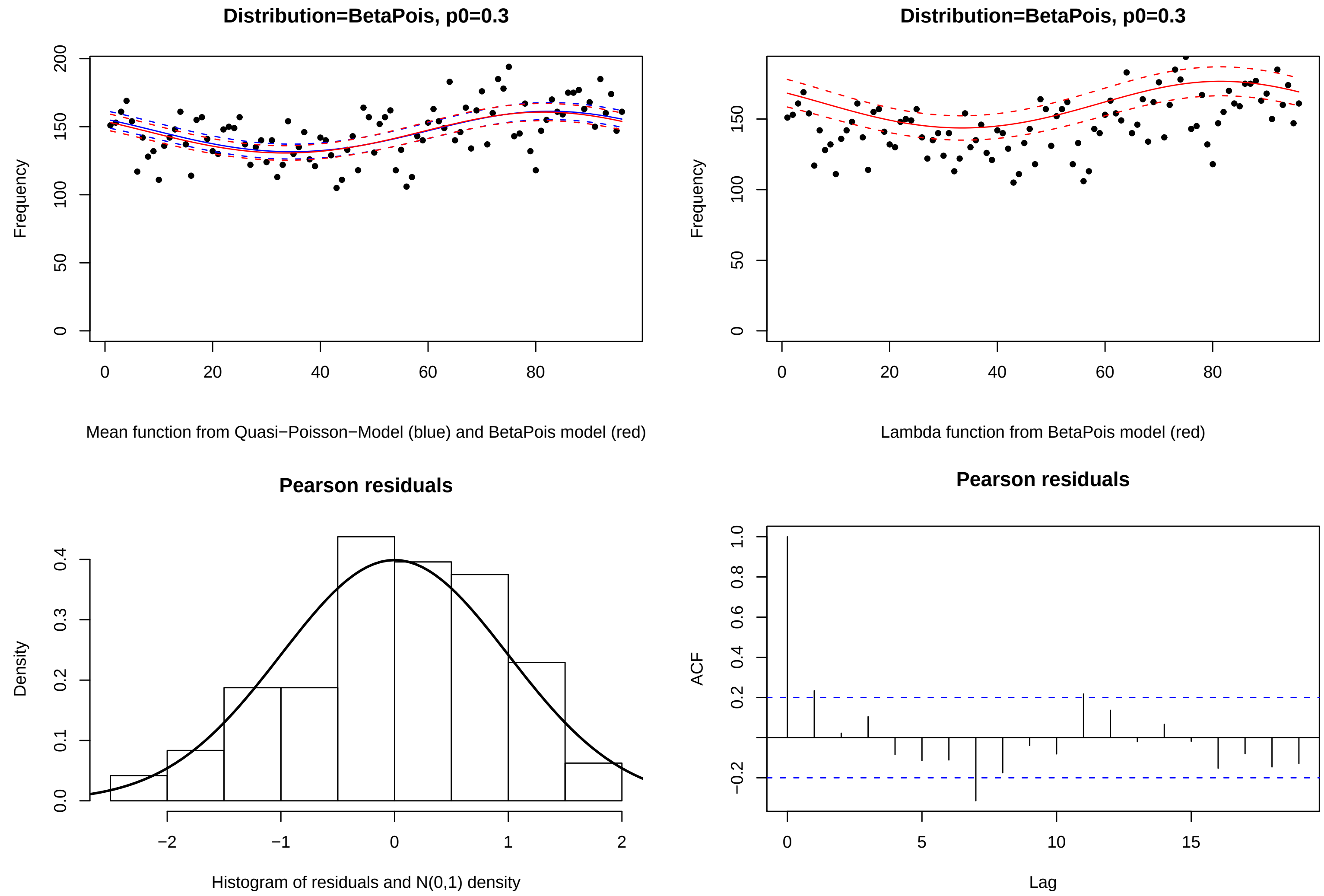
```
Distribution:  BetaPois
Formula: y ~ beta01 + T.cos1 + T.sin1 - 1

       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:
                   measures
loglik             -408.170
chi.sq               80.811
df.residual          92.000
aic                 824.341
bic                 834.599

Reporting Probabilities:
        lower estimate  upper
alpha1 0.8622   0.9097 0.9571
```

Stroke Data



$$\hat{\pi} = 0.91$$

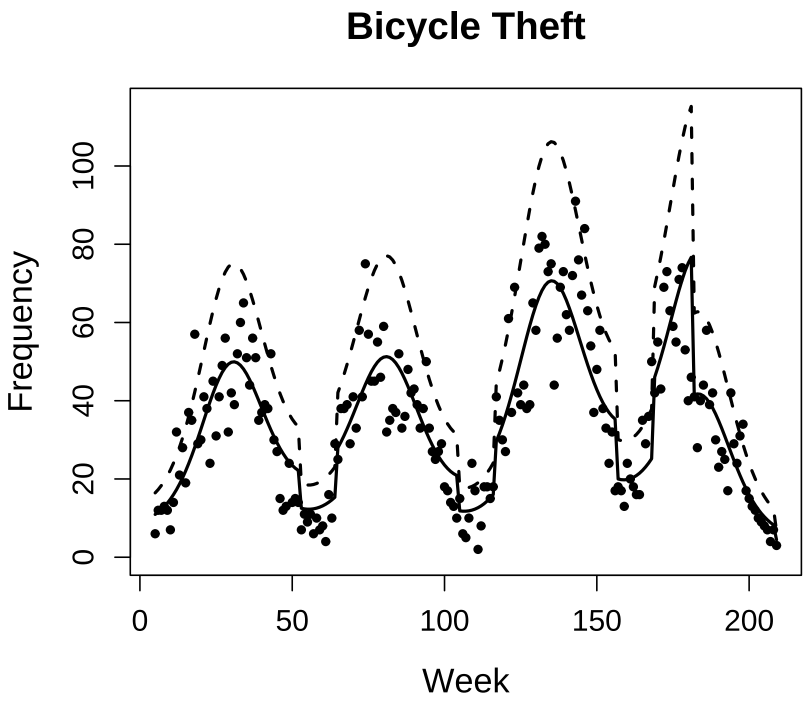
Crime Register Data

`SIMO`: Austrian online crime register

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

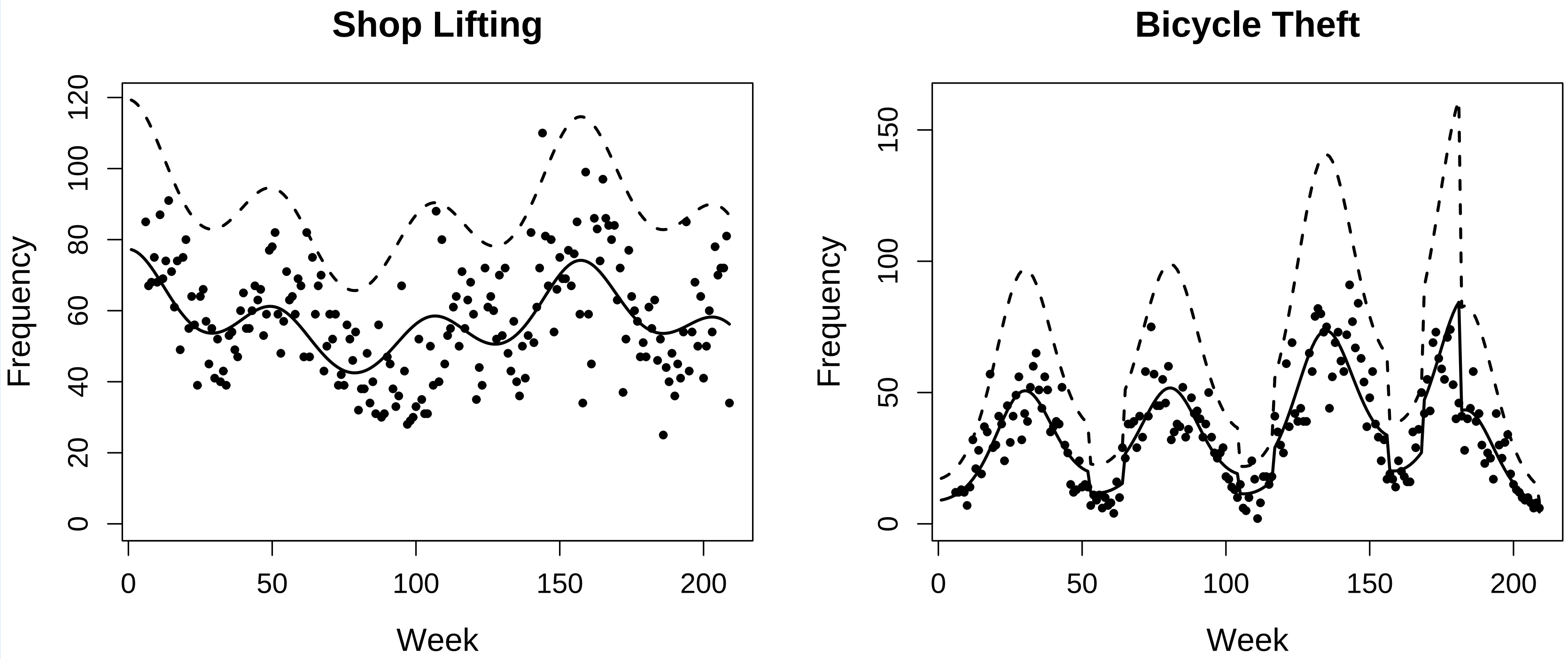
In most cases Poisson overdispersion

GP model estimates



$\hat{\pi} = 0.69$ $\hat{\pi} = 0.67$

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