Automated and Early Detection of Disease Outbreaks

Kasper Schou Telkamp

Supervisors: Jan Kloppenborg Møller, Lasse Engbo Christiansen

Master Thesis Project

Spring, 2023

Technical University of Denmark

DTU Compute

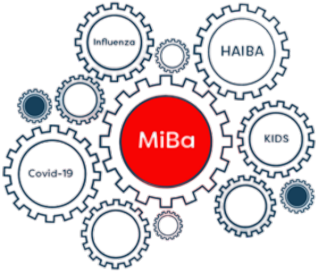
Department of Applied Mathematics and Computer Science

About me

- MSc Eng., Quantitative Biology and Disease Modelling @ DTU
- Wife and three kids @ Vedbæk
- Passionate about e-Sports, particularly Counter Strike

Motivation

- Establishment of MiBa by SSI in 2010
- Great opportunity for data analysis
- No fully automated procedures in place at SSI



State-of-the-art algorithms

State-of-the-art algorithms for aberration detection is presented in Salmon, Schumacher, and Höhle 2016 and implemented in the R package **surveillance**. The R package includes the method introduced by Farrington et al. 1996 together with the subsequently improved method proposed by Noufaily et al. 2013.

Novel algorithm

The novel algorithm utilizes a generalized mixed effects model or a hierarchical mixed effects model as a modeling framework to model the count case observations $\boldsymbol{y}$ and assess the unobserved random effects $\boldsymbol{u}$. These random effects are used directly to characterize an outbreak.

Formulation of hierarchical models

Poisson Normal

$$\boldsymbol{Y}|\boldsymbol{u} \sim \mathrm{Pois}\left(\boldsymbol{\lambda}\exp(\boldsymbol{u})\right)$$
$$\boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, I\sigma^2)$$

Poisson Gamma

$$\boldsymbol{Y}|\boldsymbol{u} \sim \mathrm{Pois}(\boldsymbol{\lambda}\boldsymbol{u})$$
$$\boldsymbol{u} \sim \mathrm{G}(\boldsymbol{1}/\phi, \phi)$$

Formulation of hierarchical models

Poisson Normal

$$\boldsymbol{Y}|\boldsymbol{u} \sim \mathrm{Pois}\left(\boldsymbol{\lambda}\exp(\boldsymbol{u})\right)$$
$$\boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, I\sigma^2)$$

Poisson Gamma

$$\boldsymbol{Y}|\boldsymbol{u} \sim \mathrm{Pois}(\boldsymbol{\lambda}\boldsymbol{u})$$
$$\boldsymbol{u} \sim \mathrm{G}(\boldsymbol{1}/\phi, \phi)$$

$$Y \sim \mathrm{NB}\left(1/\phi, 1/(\lambda\phi + 1)\right)$$

Step 1: Modeling framework

- Assume a hierarchical Poisson Normal or Poisson Gamma model to reference data using a log link
- Incorporate covariates by supplying a model formula on the form

$$\log(\lambda_{it}) = \boldsymbol{x}_{it}\boldsymbol{\beta} + \log(n_{it}), \quad i = 1, \ldots, m, \quad t = 1, \ldots, T \tag{1}$$

- Account for structural changes in the time series using a rolling window of width k

Step 2: Inference of random effects

- Infer one-step ahead random effects $\hat{u}_{it_1}$ for each group using the fitted model
- Define outbreak detection threshold U_{t_0} as a quantile of the second stage model's random effects distribution
- Use either a Gaussian or Gamma distribution with respective plug-in estimates

Step 3: Parameter estimations and outbreak detection

- Compare inferred random effects $\hat{u}_{it_1}$ to a threshold U_{t_0}
- Raise and alarm if the inferred random effect exceeds the threshold, i.e. $\hat{u}_{it_1} > U_{t_0}$
- Omit outbreak related observations from future parameter estimation

Case study

Shiga toxin (verotoxin)-producing *Escherichia coli* (STEC)

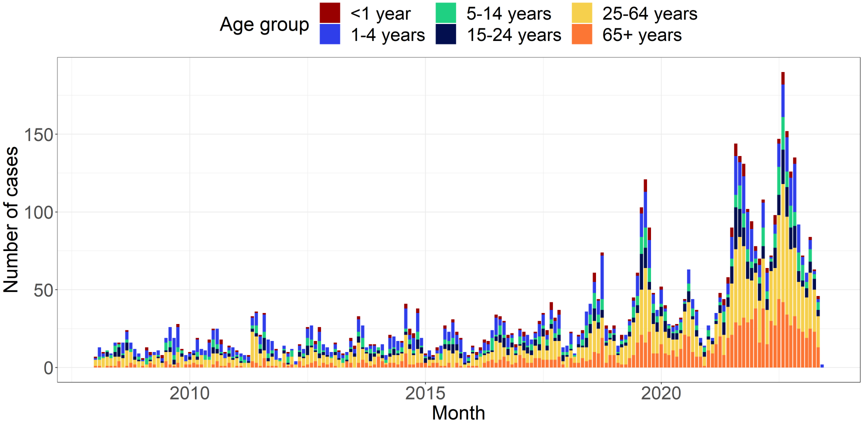


Figure: A stacked bar graph illustrating the number of monthly STEC cases observed in the period from 2008 to 2022 for the six age groups.

Constant model

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \log(n_{it}) \tag{2}$$

- λ_{it} is the outbreak intensity at time t for age group i
- $\beta(ageGroup_i)$ is the fixed effect specific to age group i
- $\log(n_{it})$ acts as an offset, accounting for the population size at time t for age group i

Trend model

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \beta_{trend}t + \log(n_{it}) \tag{3}$$

- In addition to constant model, includes a trend component
- β_{trend} quantifies the rate of change in the outbreak intensity over time

Seasonality model

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \sin\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\sin} + \cos\left(2\frac{\pi \cdot \tau_t}{12}\right)\beta_{\cos} + \log(n_{it}) \tag{4}$$

- In addition to constant model, incorporates an annual seasonality pattern
- τ_t represents the time period t within a year (1-12)
- $\beta_{\sin}$ and $\beta_{\cos}$ capture the effect of the seasonal pattern

Combined trend and seasonality model

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \beta_{trend}t + \sin\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\sin} + \cos\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\cos} + \log(n_{it}) \quad (5)$$

- Builds upon previous models, combining trend and seasonality components
- Includes both β_{trend}, $\beta_{\sin}$, and $\beta_{\cos}$ parameters

Estimated one-step ahead random effects

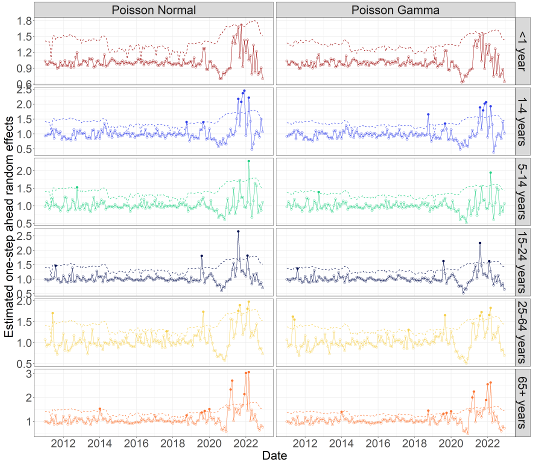
- A rolling window of width $k = 36$ months is employed
- The combined model minimizes the logarithmic score
- Upper bound U_{t_0} is based on the 90% quantile of the random effects distribution
- If the one-step ahead random effects u_{it_1} exceeds U_{t_0} an alarm is raised
- 30 alarms are generated using the Poisson Normal framework, while 31 alarms are generated using the Poisson Gamma framework.
- A great number of alarms are generated in the period from March 2021 to March 2022

Case study

Estimated one-step ahead random effects

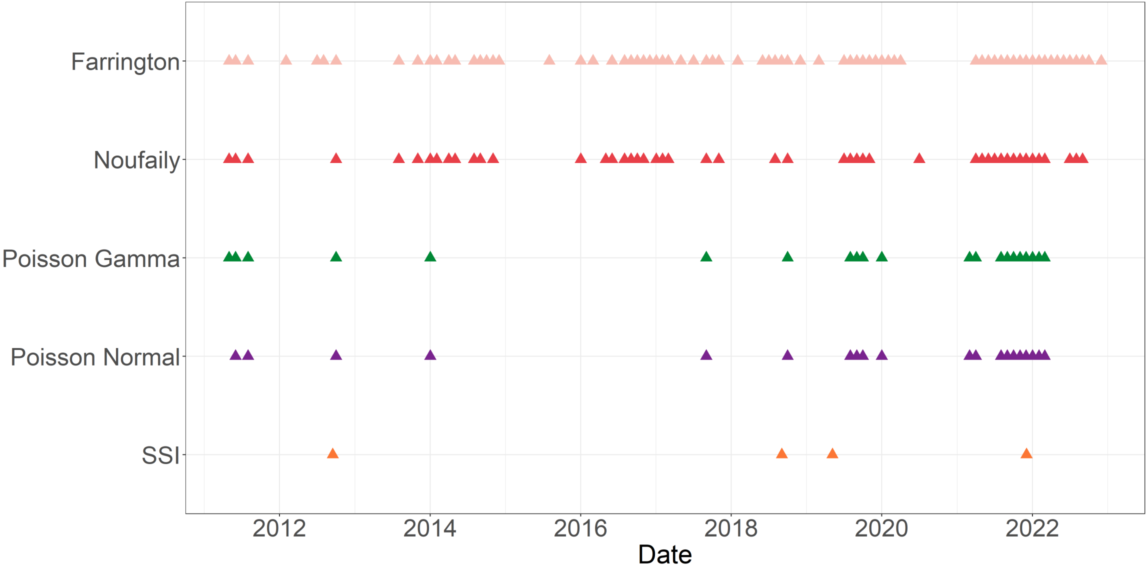
- A rolling window of width $k = 36$ months is employed
- The combined model minimizes the logarithmic score
- Upper bound U_{t_0} is based on the 90% quantile of the random effects distribution
- If the one-step ahead random effects u_{it_1} exceeds U_{t_0} an alarm is raised

- 30 alarms are generated using the Poisson Normal framework, while 31 alarms are generated using the Poisson Gamma framework.
- A great number of alarms are generated in the period from March 2021 to March 2022



Case study

Performance of statistical outbreak detection algorithms



Baseline data

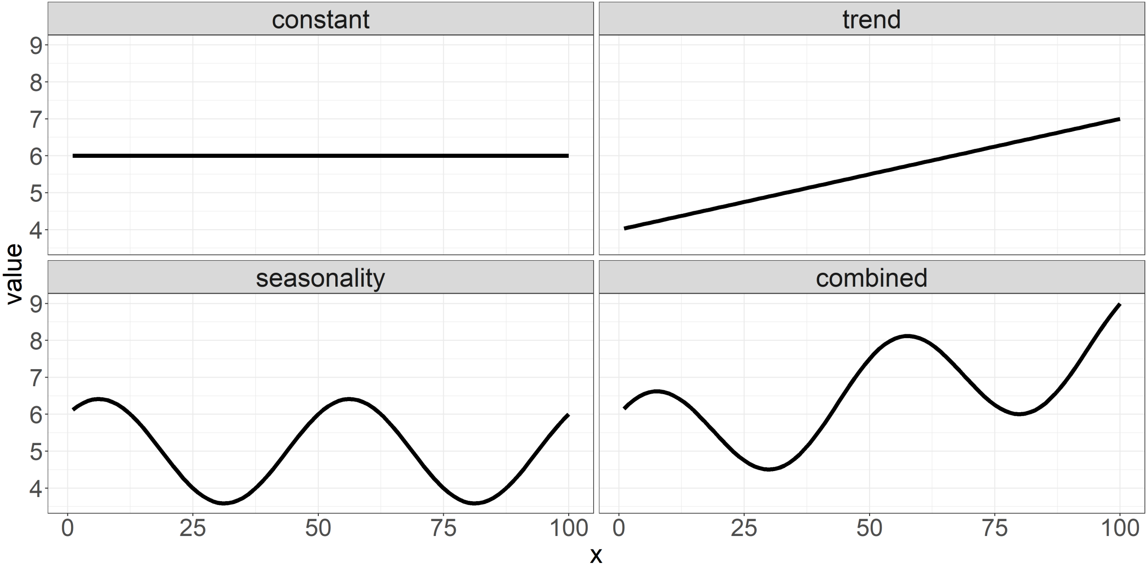
Simulated baseline data is generated according to a Negative Binomial distribution with mean μ and a variance parameter $\phi\mu$. The equation for the mean $\mu(t)$ is given as:

$$\mu(t) = \exp\left(\theta + \beta_{trend}t + \sum_{j=1}^{m}\left(\gamma_1 \cos\left(\frac{2\pi jt}{52}\right) + \gamma_2 \sin\left(\frac{2\pi jt}{52}\right)\right)\right) \quad (6)$$

Scenarios

Scenario	θ	ϕ	β	γ_1	γ_2	m	Trend
1	0.1	1.5	0.0000	0.00	0.00	0	0
2	0.1	1.5	0.0000	0.60	0.60	1	0
3	0.1	1.5	0.0025	0.00	0.00	0	1
4	0.1	1.5	0.0025	0.60	0.60	1	1
5	-2.0	2.0	0.0000	0.00	0.00	0	0
6	-2.0	2.0	0.0000	0.10	0.30	1	0
7	-2.0	2.0	0.0050	0.00	0.00	0	1
8	-2.0	2.0	0.0050	0.10	0.30	1	1
...	...	...	...	...	...	...	...
25	5.0	1.2	0.0000	0.00	0.00	0	0
26	5.0	1.2	0.0000	0.05	0.01	1	0
27	5.0	1.2	0.0001	0.00	0.00	0	1
28	5.0	1.2	0.0001	0.05	0.01	1	1

Scenarios illustration



Outbreaks

- Four outbreaks during baseline weeks (313-575), one outbreak during current weeks (576-624)
- Random constant value k is chosen
- Outbreak size v is generated from a Poisson distribution with mean equal to k times the standard deviation from the baseline data
- The v outbreak cases are distributed randomly in time according to a discretized log-normal distribution represented as $Z \sim \lfloor \mathrm{LN}(0, 0.5^2) \rfloor$

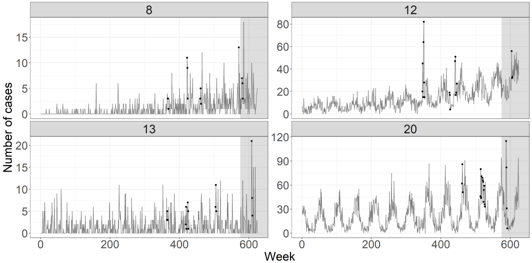
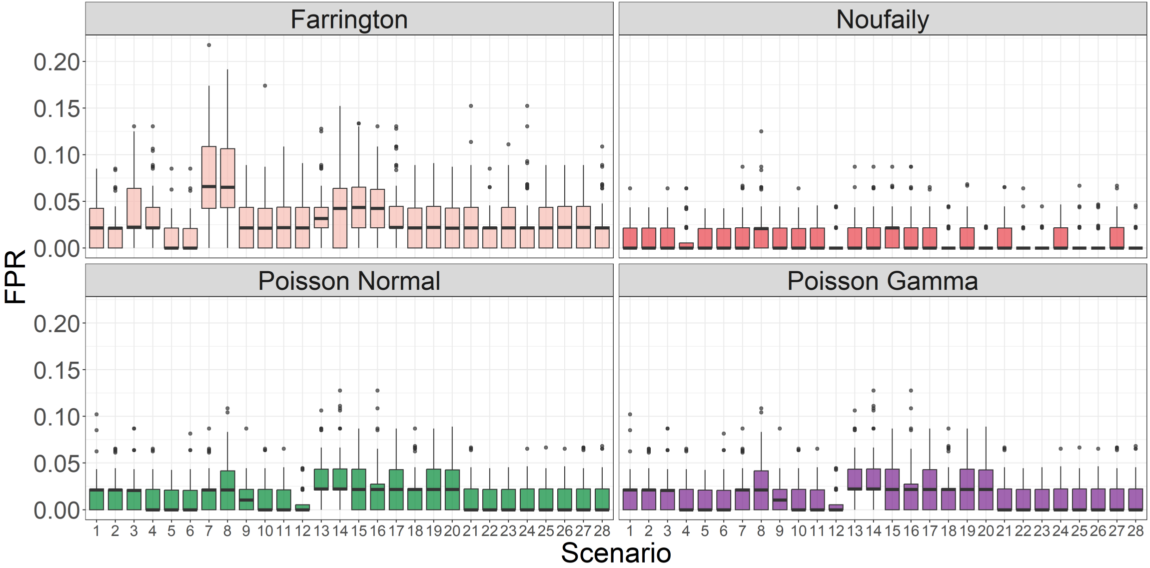
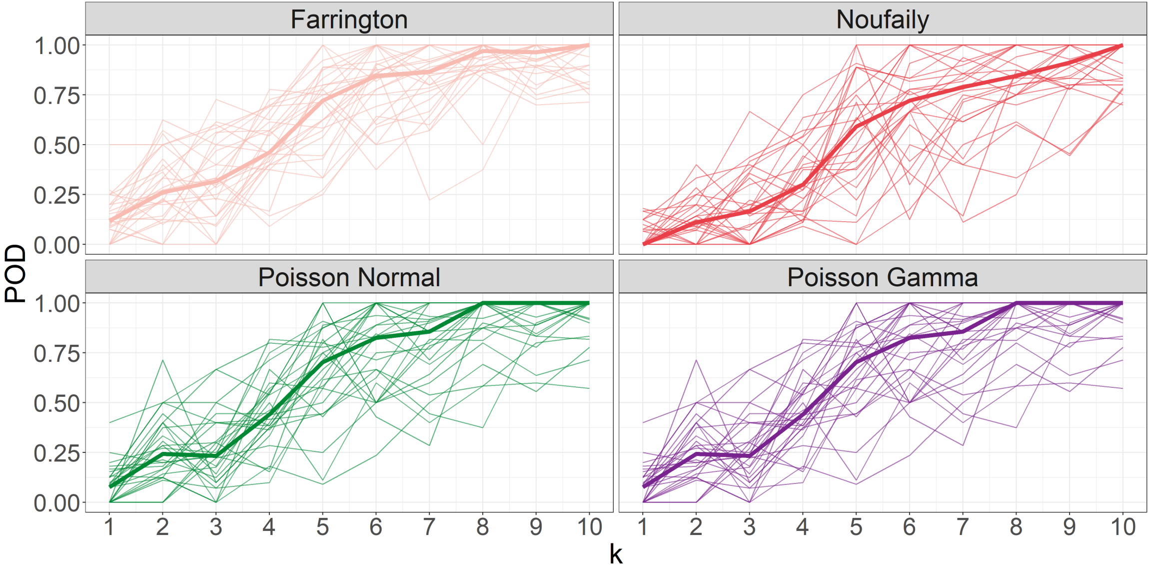


Figure: Plots of one randomly chosen realization for scenario 8, 12, 13, and 20.

False Positive Rates

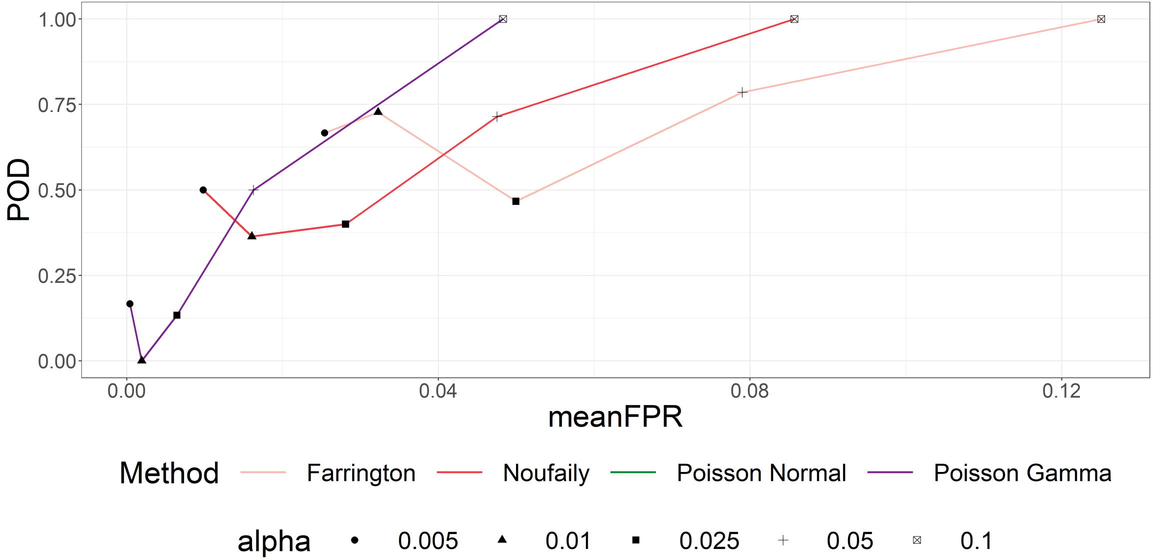


Probability an outbreak is detected



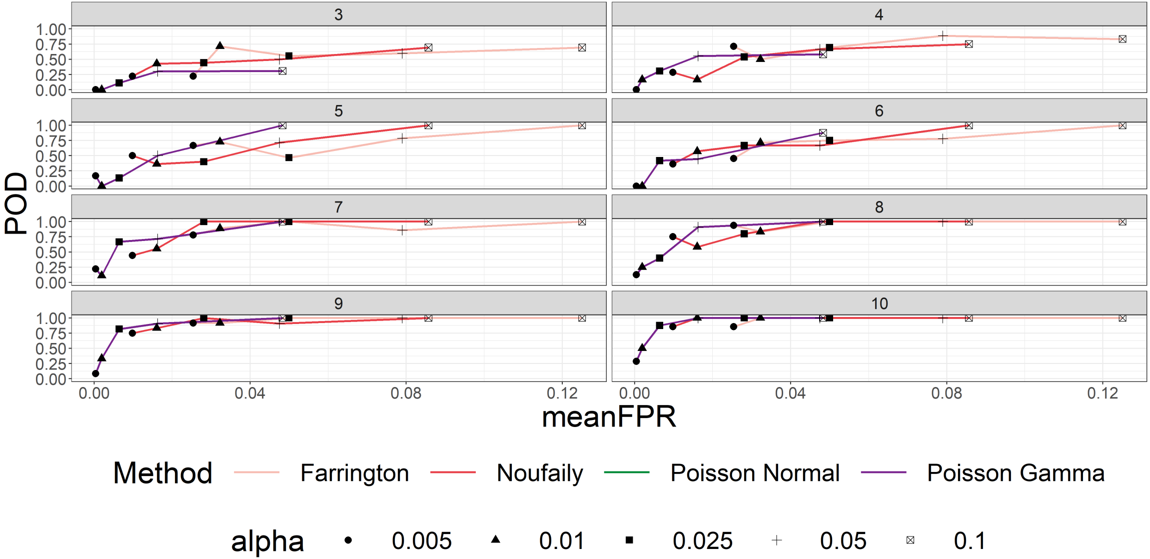
Simulation study

Profile of POD vs FPR



Simulation study

Profile of POD vs FPR in facets



Summary

- Easy incorporation of **covariates**
- Estimates are **consistent** across the two modeling frameworks
- Positively **identified outbreaks** coinciding with well-documented outbreaks
- Effectively **control the number of "false alarms"**
- Great potential in utilizing **MiBa-based surveillance**

References

Farrington, C. P. et al. (1996). "A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159.3, pp. 547–563. ISSN: 09641998, 1467985X. URL: http://www.jstor.org/stable/2983331 (visited on 01/27/2023).

Noufaily, Angela et al. (2013). "An Improved Algorithm for Outbreak Detection in Multiple Surveillance Systems". en. In: *Online Journal of Public Health Informatics* 32.7, pp. 1206–1222.

Salmon, Maëlle, Dirk Schumacher, and Michael Höhle (2016). "Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance". In: *Journal of Statistical Software* 70.10, pp. 1–35. DOI: 10.18637/jss.v070.i10. URL: https://www.jstatsoft.org/index.php/jss/article/view/v070i10.

Probability function for Y

$$
\begin{aligned}
P[Y = y] &= g_Y(y; \boldsymbol{\beta}, \phi) \\
&= \frac{\lambda^y}{y!\Gamma(1/\phi)\phi^{1/\phi}} \frac{\phi^{y+1/\phi}\Gamma(y+1/\phi)}{(\lambda\phi+1)^{y+1/\phi}} \\
&= \frac{\Gamma(y+1/\phi)}{\Gamma(1/\phi)y!} \frac{1}{(\lambda\phi+1)^{1/\phi}} \left(\frac{\lambda\phi}{\lambda\phi+1}\right)^y \\
&= \binom{y+1/\phi-1}{y} \frac{1}{(\lambda\phi+1)^{1/\phi}} \left(\frac{\lambda\phi}{\lambda\phi+1}\right)^y, \quad \text{for } y = 0, 1, 2, \ldots
\end{aligned}
\tag{7}
$$

where the following convention is used

$$
\binom{z}{y} = \frac{\Gamma(z+1)}{\Gamma(z+1-y)y!} \tag{8}
$$

The marginal distribution of Y is a negative binomial distribution, $Y \sim \mathrm{NB}\left(1/\phi, 1/(\lambda\phi+1)\right)$

Proof

The probability function for the conditional distribution of Y for given u

$$f_{Y|u}(y; u, \boldsymbol{\beta}) = \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \tag{9}$$

and the probability density function for the distribution of u is

$$f_u(u; \phi) = \frac{1}{\phi \Gamma(1/\phi)} \left(\frac{u}{\phi} \right)^{1/\phi - 1} \exp(-u/\phi) \tag{10}$$

Proof

Given (9) and (10), the probability function for the marginal distribution of Y is determined from

$$\begin{aligned} g_Y(y;\beta,\phi) &= \int_{u=0}^{\infty} f_{Y|u}(y;u,\beta) f_u(u;\phi)\,du \\ &= \int_{u=0}^{\infty} \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \frac{1}{\phi\Gamma(1/\phi)} \left(\frac{u}{\phi}\right)^{1/\phi-1} \exp(-u/\phi)\,du \\ &= \frac{\lambda^y}{y!\Gamma(1/\phi)\phi^{1/\phi}} \int_{u=0}^{\infty} u^{y+1/\phi-1} \exp\big(-u(\lambda\phi+1)/\phi\big)\,du \end{aligned} \tag{11}$$

Proof

In (11) it is noted that the integrand is the *kernel* in the probability density function for a Gamma distribution, $\mathrm{G}\left(y + 1/\phi, \phi/(\lambda\phi + 1)\right)$. As the integral of the density shall equal one, we find by adjusting the norming constant that

$$\int_{u=0}^{\infty} u^{y+1/\phi-1} \exp\Big(- u/(\phi/(\lambda\phi + 1))\Big)\, du = \frac{\phi^{y+1/\phi}\Gamma(y + \mathbf{1}/\phi)}{(\lambda\phi + 1)^{y+1/\phi}} \tag{12}$$

and then (7) follows