

**A Piecewise Linear Generalized
Poisson Regression Approach to
Modeling Longitudinal Frequency
Data**

Jennifer Jo Borgesi

Advisor: John C. Kern II

Department of Mathematics and Computer
Science

Duquesne University

April 15, 2004

Outline

1. Introduction
2. Generalized Poisson Distribution
3. Generalized Poisson Application
4. Hot Flush Application
5. Discussion
6. Limitations
7. Future Work

Introduction

Problem: Inadequate modeling techniques for data from experiments that produce longitudinal frequency data.

Current Status: Different models have been fitted to such data (i.e. the Negative Binomial Distribution). These models are inadequate because they only allow for overdispersion.

Proposed Methodology: Use the Generalized Poisson Distribution to model data.

Generalized Poisson Distribution

The generalized Poisson distribution (see Consul and Jain 1973) is defined by the mass function

$$p(x|\theta, \lambda) = \theta(\theta + x\lambda)^{x-1} e^{-(\theta+x\lambda)} \frac{1}{x!}, \quad x = 0, 1, 2, \dots$$

for $\theta > 0$, and $|\lambda| < 1$, such that

$$p(x|\theta, \lambda) = 0 \quad \text{for } x \geq m \quad \text{when } \lambda < 0;$$

m is the largest positive integer for which

$$\theta + m\lambda \leq 0.$$

Generalized Poisson Distribution

It can be shown that if X is a random variable with this generalized Poisson distribution then the expected value μ of X is

$$\mu = \frac{\theta}{1 - \lambda}$$

with variance

$$\sigma^2 = \frac{\theta}{(1 - \lambda)^3}.$$

An alternative, more convenient parameterization of the generalized Poisson distribution is

$$X \sim GP(\mu, k),$$

where $\mu > 0$ is the expected value of the generalized Poisson random variable and k is the dispersion parameter (Famoye and Wang 1997).

Generalized Poisson Distribution

The GP density $f(x|\mu, k)$ in terms of the parameters μ and k is then

$$f(x|\mu, k) = \left(\frac{\mu}{1 + k\mu} \right)^x \frac{(1 + kx)^{x-1}}{x!} \\ * \exp \left(\frac{-\mu(1 + kx)}{1 + k\mu} \right); \quad x = 0, 1, 2, \dots$$

Generalized Poisson Distribution

The variance of X , denoted by VX , is given by

$$VX = \mu(1 + k\mu)^2 .$$

- for $k > 0$ the variance of X exceeds its expected value (overdispersion)
- for $\frac{-2}{\mu} < k < 0$ the expected value μ exceeds the variance of X (underdispersion)
- for $k = 0$, $\mu = VX$, and the generalized Poisson distribution reduces to a standard Poisson distribution (equidispersion)

Generalized Poisson Application

Univariate Parameter Estimation - begin by updating μ and k as follows:

- Choose current (initial) values of the parameters, call these values $\mu^{(0)}$ and $k^{(0)}$.
- Propose a value for μ^* from a uniform distribution centered at $\mu^{(0)}$ the current value.

$$\mu^* \sim \text{Unif}(\mu^{(0)} - a, \mu^{(0)} + a),$$

where $a > 0$ is a tuning parameter in the proposal distribution for μ^* . We have let the proposal distribution for μ^* be uniform which gives

$$q(\mu^* | \mu) = \frac{1}{\min(2a, \mu + a, \frac{1}{-2k})}.$$

Generalized Poisson Application

- For fixed k , the proposed value μ^* is then accepted with the following probability:

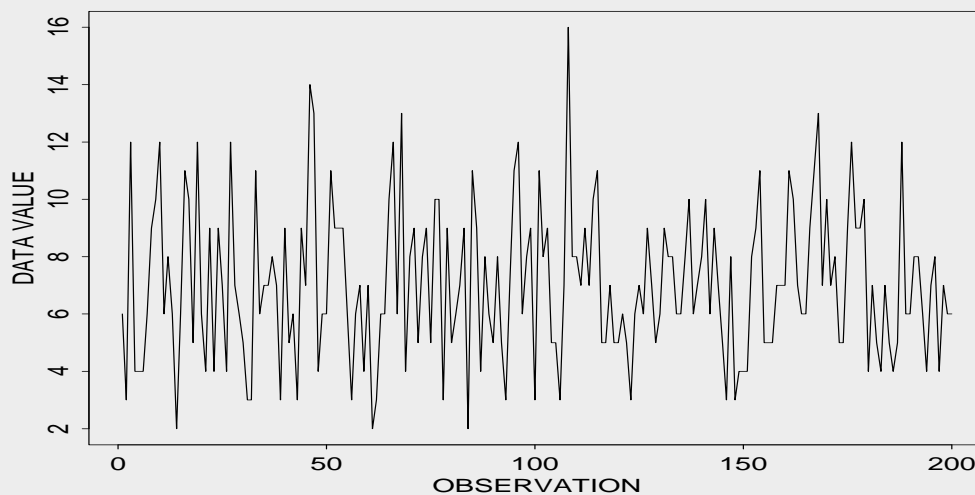
$$\alpha = \min \left(1, \frac{L(\mu^*, k)\pi(\mu^*, k)q(\mu^{(0)}|\mu^*)}{L(\mu^{(0)}, k)\pi(\mu^{(0)}, k)q(\mu^*|\mu^{(0)})} \right).$$

The value that is accepted for μ — with probability α — is then used to update k in analogous fashion.

The values $\mu^{(1)}$ and $k^{(1)}$ represent the updated values of these parameters. These values are then treated as the new “current” parameter values, and the entire updating process is repeated.

Generalized Poisson Application

The Equidispersed Data Set:

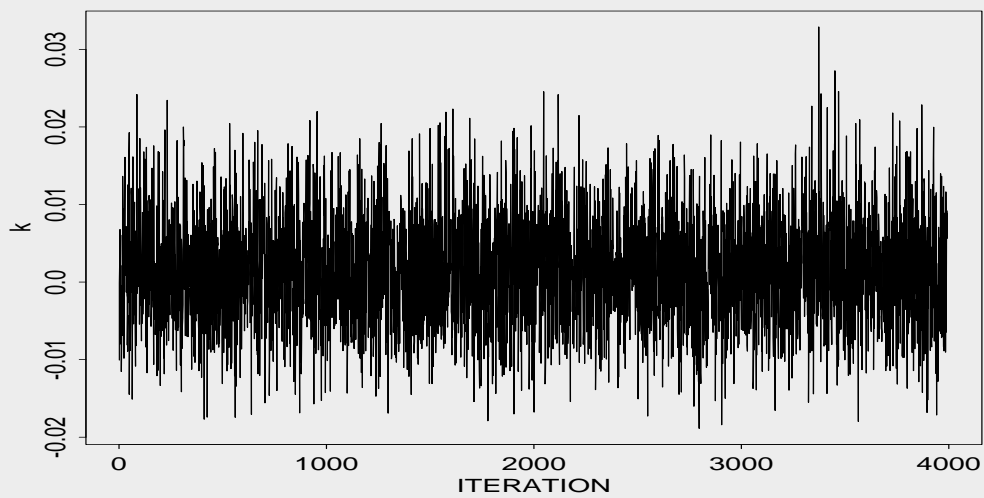
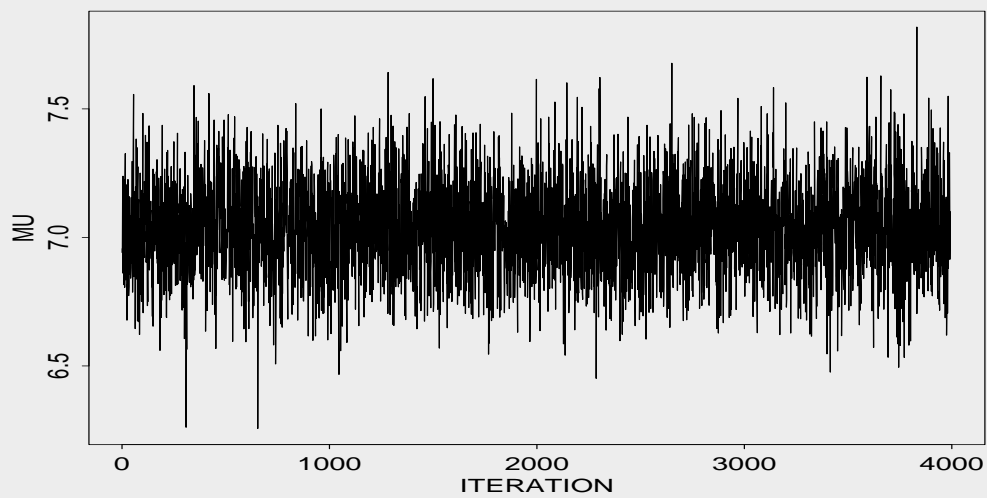


For the computer simulated equidispersed data set of $n = 200$ realizations:

- $\bar{x} = 7.03$
- $s^2 = 7.074472$
- $\hat{k} = 0.0004492$

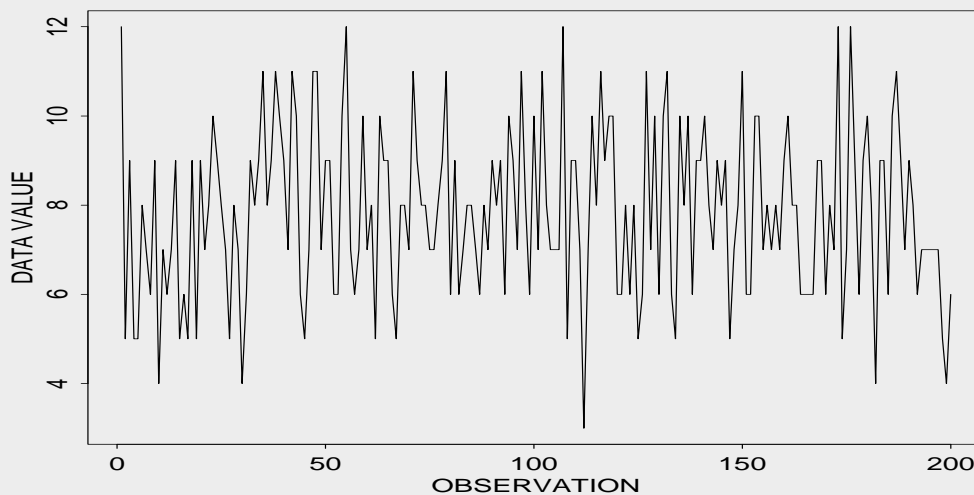
Generalized Poisson Application

The Equidispersed Data Set:



Generalized Poisson Application

The Underdispersed Data Set:

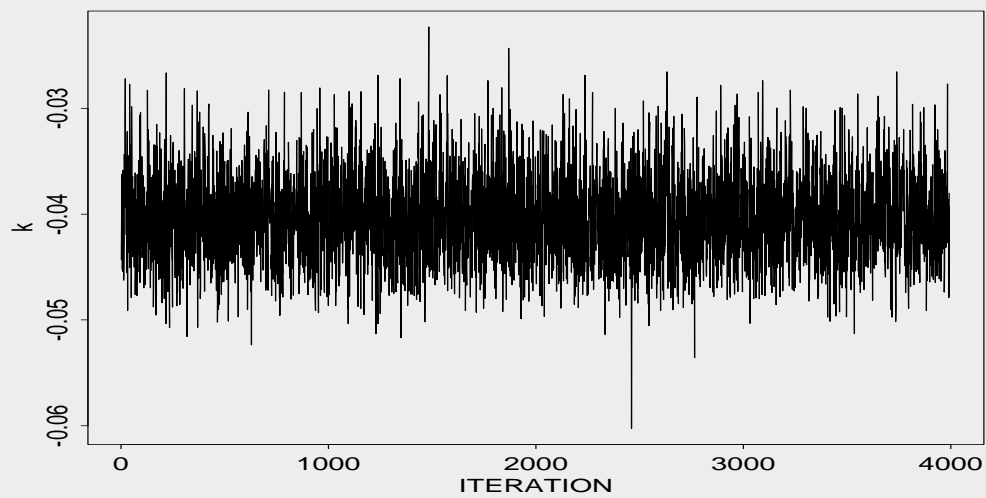
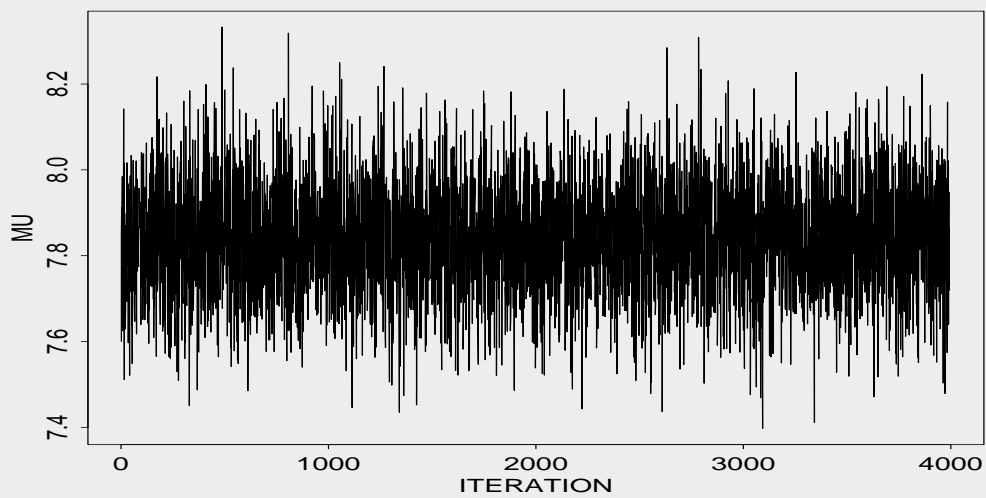


For the computer simulated underdispersed data set of $n = 200$ realizations:

- $\bar{x} = 7.84$
- $s^2 = 3.622513$
- $\hat{k} = -0.0408486677$

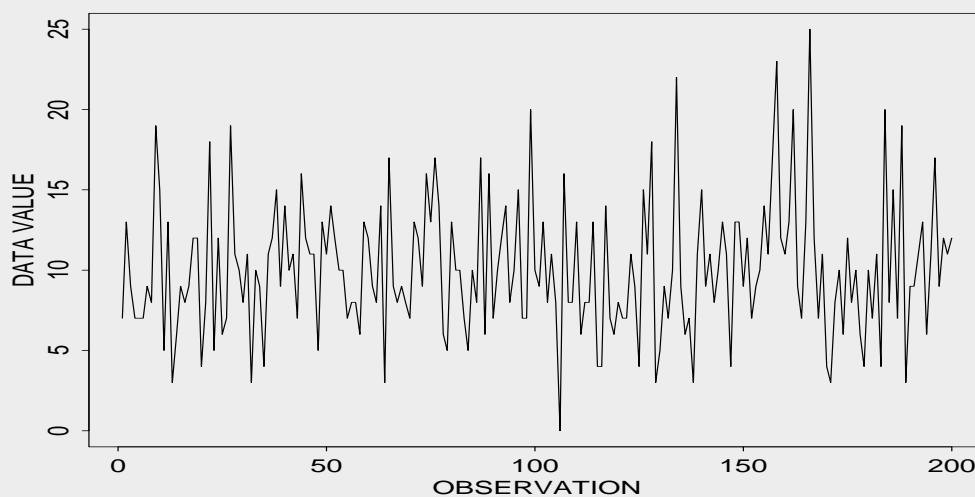
Generalized Poisson Application

The Underdispersed Data Set:



Generalized Poisson Application

The Overdispersed Data Set:

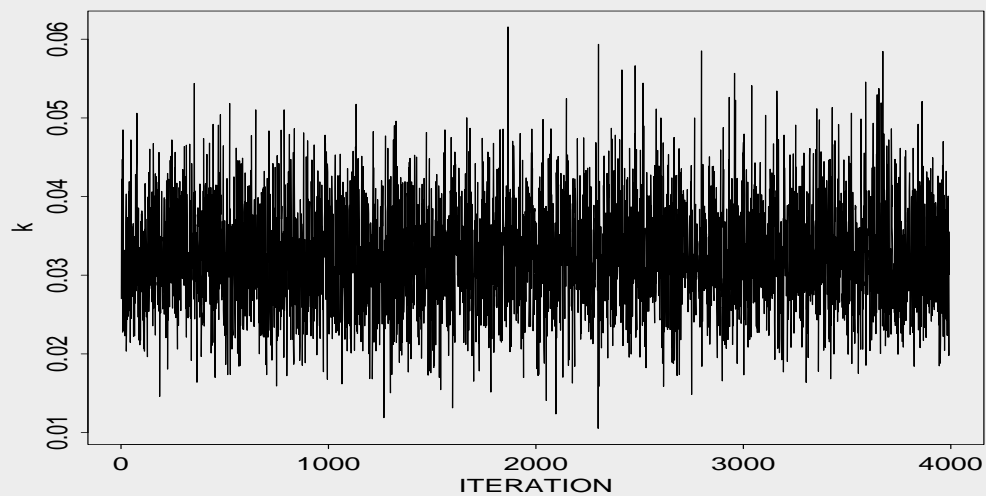
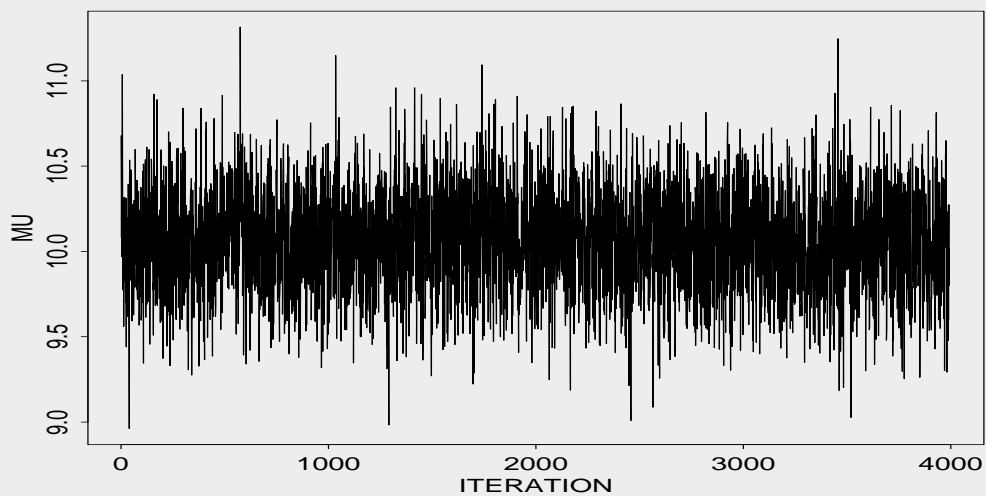


For the computer simulated overdispersed data set of $n = 200$ realizations:

- $\bar{x} = 10.05$
- $s^2 = 17.50503$
- $\hat{k} = 0.03181794$

Generalized Poisson Application

The Overdispersed Data Set:



Hot Flush Application

Data: A clinical trial investigating acupuncture as an alternative treatment to alleviate symptoms of menopause for breast cancer survivors.

- Treatment group
- Placebo group
- Education group

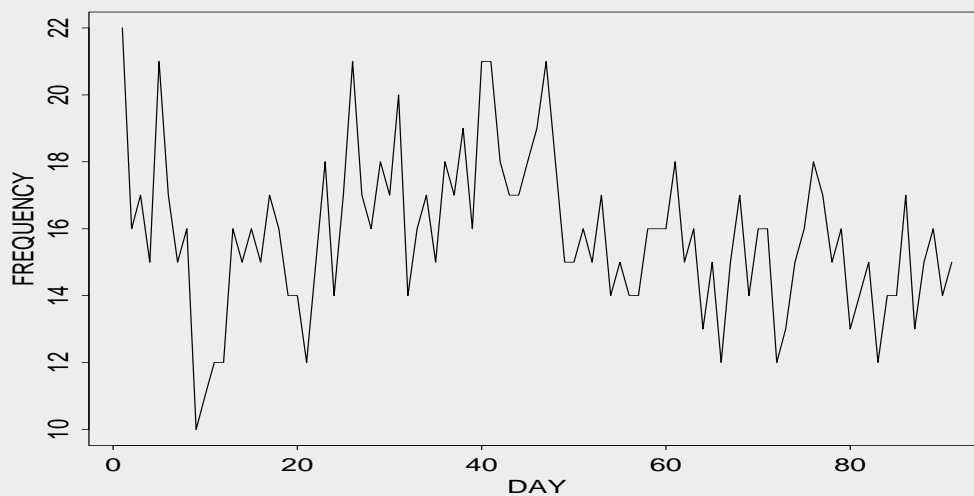


Figure 1: The HFF profile of an individual from the treatment group.

Hot Flush Application

Data Model:

We model longitudinal frequency responses from multiple individuals using a GP distribution whose mean μ is a function of time.

The log-likelihood function $l_t(\mu_t, k)$ based on the n_t observations collected at time t :

$$l_t(\mu_t, k) = \log \prod_{j=1}^{n_t} \left(\frac{\mu_t}{1 + k\mu_t} \right)^{y_{tj}} \frac{(1 + ky_{tj})^{y_{tj}-1}}{y_{tj}!} \exp \left(\frac{-\mu_t(1 + ky_{tj})}{1 + k\mu_t} \right).$$

Hot Flush Application

Flexibility in allowing μ_t to vary with time is obtained by modeling μ_t as a **piecewise-linear function** of time.

We define a vector of knot locations

$$\mathbf{K} = \{K_1, K_2, \dots, K_m\}$$

and a vector of corresponding heights

$$\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$$

(K_i, λ_i) represents the location at which two line segments meet (referred to as a *node*).

$$\mu_t = f(\lambda_i, K_i, t) = \left(\frac{\lambda_{i+1} - \lambda_i}{K_{i+1} - K_i} \right) (t - K_i) + \lambda_i,$$

for $i = 1, \dots, m$.

Hot Flush Application

Five nodes were selected to formulate the piecewise linear function.

- three fixed knot locations $\{0.5, 7.5, 91.5\}$
- two additional knot locations randomly chosen in the time interval $(7.5, 91.5)$

$$\mathbf{K} = \{K_1, K_2, \dots, K_5\}$$

$$\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_5\}$$

$$K_1 < K_2 < K_3 < K_4 < K_5 \text{ and } \lambda_i > 0.$$

The complete list of parameters that defines this model is $\{\mathbf{K}, \boldsymbol{\lambda}, k\}$, where k is the dispersion term.

Hot Flush Application: Parameter Estimation

Let

$$\{\lambda_1^{(i)}, \dots, \lambda_5^{(i)}, K_3^{(i)}, K_4^{(i)}, k^{(i)}\}$$

represent the “current” value of the parameters.

- To update λ_1 let the proposal distribution be uniform over an interval centered around the current value $\lambda_1^{(i)}$. Then

$$\lambda_1^* \sim \text{Unif}(\lambda_1^{(i)} - a, \lambda_1^{(i)} + a)$$

where a is a tuning parameter. For fixed values of the other 7 parameters, accept λ_1^* as the next current value with probability α given by

$$\alpha = \min \left(1, \frac{L(\lambda_1^*, \lambda_2^{(i)}, \dots, \lambda_5^{(i)}, \mathbf{K}^{(i)}, k^{(i)}) \pi(\lambda_1^*) q(\lambda_1^{(i)} | \lambda_1^*)}{L(\lambda_1^{(i)}, \lambda_2^{(i)}, \dots, \lambda_5^{(i)}, \mathbf{K}^{(i)}, k^{(i)}) \pi(\lambda_1^{(i)}) q(\lambda_1^* | \lambda_1^{(i)})} \right).$$

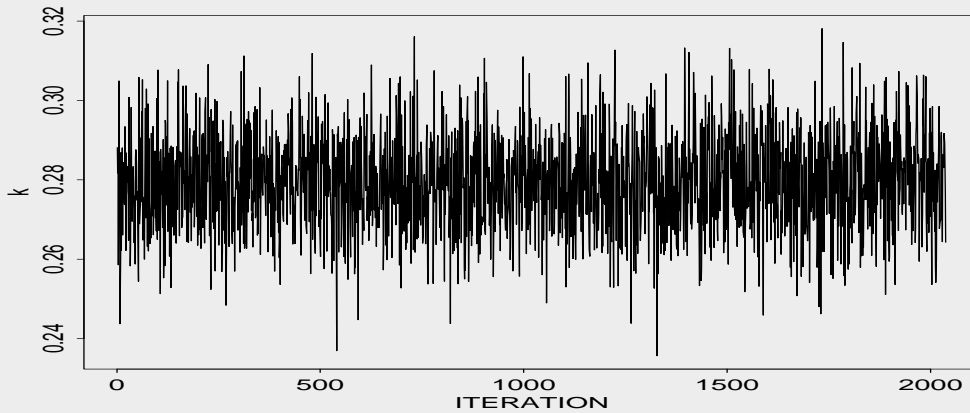
The other seven parameters are then updated using the “current” value of λ_1 .

Hot Flush Application: Parameter Estimation

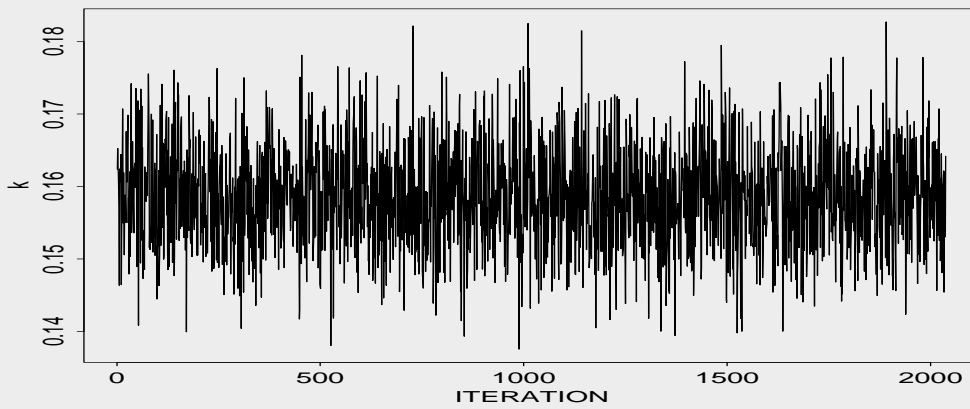
- The remaining node heights, $\lambda_2, \dots, \lambda_5$, are updated in a manner analogous to updating λ_1 .
- K_3 follows the same process as updating λ_1 , but with the following exception: the proposed value of K_3^* must be discrete uniform.
- K_4 is updated in a manner that is analogous to updating K_3
- The dispersion parameter, k , is updated in the same way as λ_1 .

After this cycle of updating the 8 parameters in turn is complete, it is repeated. This process is continued until the values stored for $\{\boldsymbol{\lambda}^{(i+1)}, \mathbf{K}^{(i+1)}, k^{(i+1)}\}$ have converged to the posterior distribution for $\{\boldsymbol{\lambda}, \mathbf{K}, k\}$.

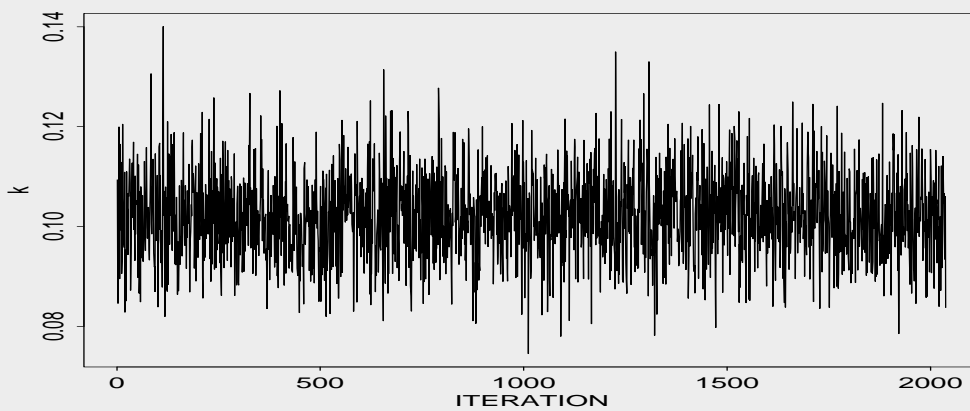
Hot Flush Application: Results



Trace plot of the marginal posterior draws for k in the treatment group.

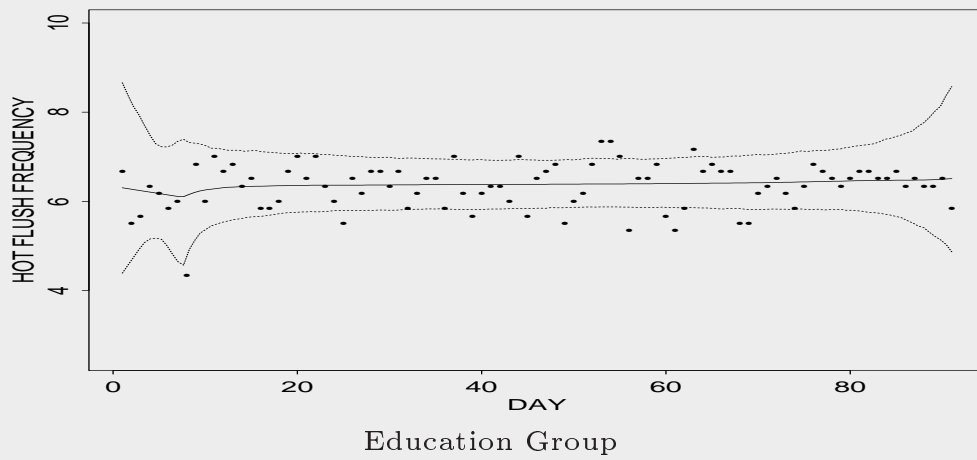
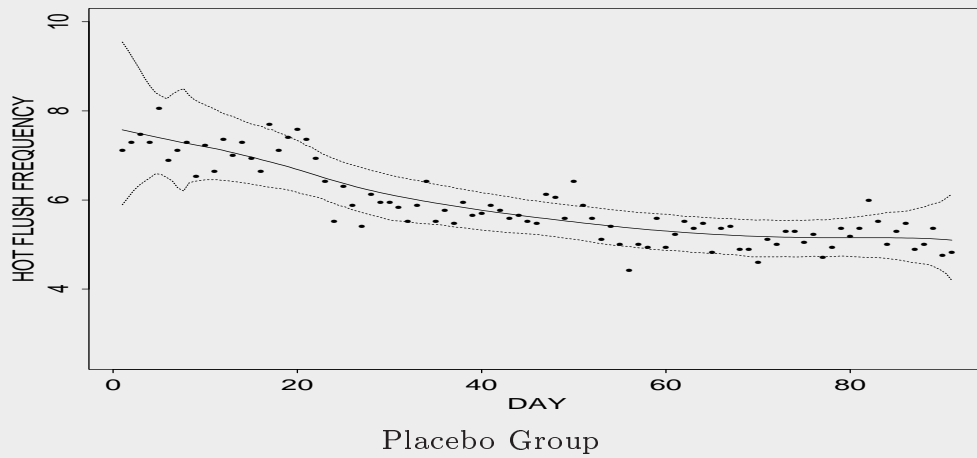
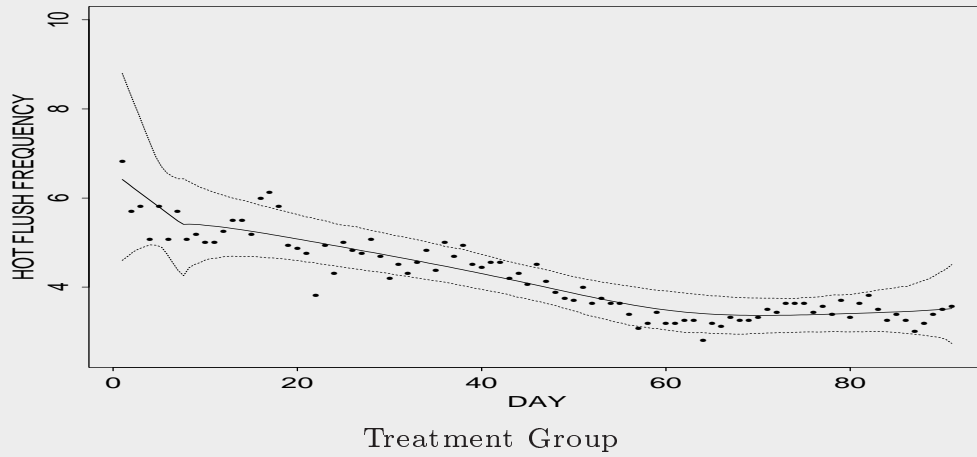


Trace plot of the marginal posterior draws for k in the placebo group.



Trace plot of the marginal posterior draws for k in the education group.

Hot Flush Application: Results



Discussion

Usefulness of the GP distribution:

- Recognizes and treats the discrete nature of the data.
- Not only allows for, but detects equidispersed, underdispersed, and overdispersed data.

Usefulness of the piecewise-linear model:

- Ability to make comparisons across experimental groups.
- Allows for time correlation through the piecewise linear function for the mean.
- Useful in other applications involving longitudinal frequency data.

Limitations

The value chosen for k , determines the range of values that μ can take:

$$k = \frac{\lambda}{\theta}, \mu = \frac{\theta}{1-\lambda}, \text{ and } -1 < \lambda < 0.$$

By substitution,

$$\begin{aligned} -1 < k\theta < 0, \\ 0 < \theta < -k. \end{aligned}$$

$\mu = \frac{\theta}{1-k\theta}$ over the interval for θ gives

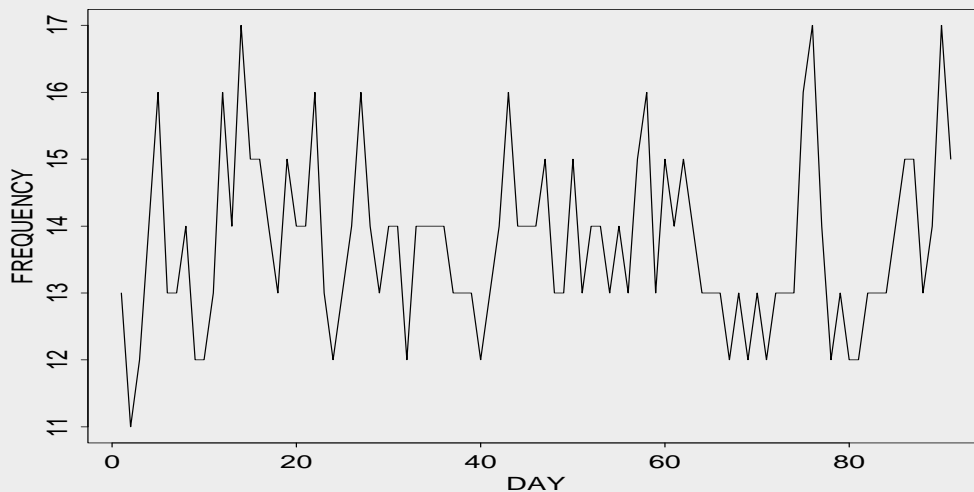
$$0 < \mu < \frac{1}{-2k}$$

and

$$\frac{-1}{2\mu} < k < 0.$$

Limitations

Example: Daily hot flush frequencies experienced by a subject in the placebo group.



- $\bar{x} = 13.73626$
- $s^2 = 1.640781$
- $\hat{k} = -0.0478$

Upper bound on μ :

$$\frac{1}{-2k} = \frac{1}{0.0956} = 10.46025.$$

Future Work

- Assign a mixture prior distribution for k (i.e. allow $\pi(k = 0) \geq 0$).
- Compare with alternative longitudinal frequency models (i.e. a model that explicitly incorporates the dependence structure of the data).