Spring 2004

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

Jennifer Borgesi

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

A Thesis

Presented to the Faculty

of the Department of Mathematics and Computer Science

McAnulty College and Graduate School of Liberal Arts

Duquesne University

in partial fulfillment of

the requirements for the degree of

Masters of Science in Computational Mathematics

by

Jennifer Jo Borgesi

April 16, 2004

Jennifer Jo Borgesi

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

Master of Science in Computational Mathematics

Department of Mathematics & Computer Science
Duquesne University, Pittsburgh, PA, USA

April 16, 2004

APPROVED_____
John Kern, Ph.D., Assistant Professor
Department of Mathematics & Computer Science

APPROVED_____
Frank D'Amico, Ph.D., Chair
Department of Mathematics & Computer Science

APPROVED_____
Kathleen Taylor, Ph.D.,
Graduate Director of Computational Mathematics
Department of Mathematics & Computer Science

APPROVED_____
Constance D. Ramirez, Ph.D., Dean
McAnulty College and Graduate School of Liberal Arts

Acknowledgments

I would like to thank my adviser Dr. John Kern for his unending support and guidance. He has the ability to bring out a person's fullest potential and I owe him my wholehearted gratitude.

A special thanks goes out to my committee and all of the Math and Computer Science faculty and staff. I only wish that I could put into words how wonderful they made my experience here at Duquesne.

I would also like to thank my family for the love and happiness they bring into my life and my fellow graduate students, especially Joe, Melissa, and Sara, for helping me survive throughout this entire process.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem to Investigate

In this research we consider experiments that generate longitudinal frequency data. Quite often this data comes from two or more experimental groups. Experiments that yield such data are common in the medical field and are often designed with the purpose of ascertaining differences among experimental groups. Standard modeling techniques, such as Repeated Measures Anova, are inadequate for expressing longitudinal frequency data because they ignore the correlation between the measurements and the discrete nature of the data. The objective of this thesis is to create a statistical model capable of sufficiently representing longitudinal frequency data.

1.2 Present Status and Proposed Methodology

Different models have been fitted to longitudinal frequency data including those that assume the data to have come from a Negative Binomial Distribution (Kern and Cohen 2003). This distribution, however, allows only for overdispersed data (where the variance exceeds the mean), thus limiting the model's applicability. We propose an alternative distribution—the generalized Poisson (GP) distribution—which allows

for both underdispersed and overdispersed data, thus making the model more flexible. This distribution has seen application in many areas; Famoye 1993 examines the GP distribution in a regression context, and Famoye and Wang 1997 use this distribution to model household fertility decisions. In this research, we apply the GP distribution to simulated data, as well as to real data collected from menopausal women. Before defining these models, we present the details of the GP distribution.

1.3 Generalized Poisson Distribution

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

$$p(x|\theta,\lambda) \;=\; \theta(\theta+x\lambda)^{x-1}e^{-(\theta+x\lambda)}\frac{1}{x!}, \quad x \;=\; 0,1,2,\ldots \tag{1.1}$$

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

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

m is the largest positive integer for which $\theta + m\lambda \;\leq 0$.

The GP distribution, like all probability distributions, must assign nonnegative probabilities that sum to one. Confirming the nonnegative requirement is straightforward from casual inspection of the mass function and its parameter constraints. Showing that the GP probabilities sum to one (i.e. $\sum_x p(x|\theta,\lambda) = 1$) requires slightly more effort. To do this, we use Lagrange's Expansion (Weisstein 1999):

$$\phi(z) = \phi(0) + \sum_{1}^{\infty}\left(\frac{d^{x-1}}{dz^{x-1}}(f(z))^x\phi'(z)\right)_{z=0}\left(\frac{z}{f(z)}\right)^x\frac{1}{x!}\,,$$

and set

$$\phi(z) = e^{\theta z} \text{ and } f(z) = e^{\lambda z} \ .$$

Substitution yields

$$\begin{aligned} e^{\theta z} &= e^0 + \sum_1^\infty \left(\frac{d^{x-1}}{dz^{x-1}} \left[e^{\lambda z x} \frac{d}{dz}(e^{\theta z}) \right] \right)_{z=0} \left(\frac{z}{e^{\lambda z}} \right)^x \\ &= 1 + \sum_1^\infty \left(\theta(\theta + \lambda x)^{x-1} e^{z(\lambda x + \theta)} \right)_{z=0} z^x e^{-\lambda x z} \frac{1}{x!} \\ &= 1 + \sum_1^\infty \theta(\theta + \lambda x)^{x-1} z^x e^{-\lambda x z} \frac{1}{x!} \\ &= \sum_{x=0}^\infty \theta(\theta + x\lambda)^{x-1} z^x e^{-\lambda x z} \frac{1}{x!} \ . \end{aligned}$$

From this last expression, we show that $\sum p(x|\theta, \lambda) = 1$.

$$\begin{aligned} e^{\theta z} &= \sum_{x=0}^\infty \theta(\theta + x\lambda)^{x-1} z^x e^{-\lambda x z} \frac{1}{x!} \\ \frac{e^{\theta z}}{e^{\theta z}} &= \frac{\sum_{x=0}^\infty \theta(\theta + x\lambda)^{x-1} z^x e^{-\lambda x z} \frac{1}{x!}}{e^{\theta z}} \\ 1 &= \sum_{x=0}^\infty \theta(\theta + x\lambda)^{x-1} z^x e^{-\lambda x z - \theta z} \frac{1}{x!} \\ 1 &= \sum_{x=0}^\infty \theta(\theta + x\lambda)^{x-1} z^x e^{-z(\lambda x + \theta)} \frac{1}{x!} \ . \end{aligned}$$

Now setting $z = 1$ gives

$$1 = \sum_{x=0}^\infty \theta(\theta + x\lambda)^{x-1} e^{-(\lambda x + \theta)} \frac{1}{x!},$$

and the proof of $\sum_x p(x|\theta, \lambda) = 1$ is complete. It can be shown that if X is a random variable with this GP 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 GP distribution is

$$X \sim GP(\mu, k), \tag{1.2}$$

where $\mu > 0$ is the expected value of the GP random variable and k is the dispersion parameter (Famoye and Wang 1997). The GP density $f(x|\mu, k)$ in terms of the parameters μ and k is then

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

and zero otherwise. Specifically, the variance of X, denoted by VX, is given by

$$VX \;=\; \mu(1+k\mu)^2\,. \tag{1.3}$$

So for $k > 0$ the variance of X exceeds its expected value μ (overdispersion); for $\frac{-2}{\mu} < k < 0$ the expected value μ exceeds the variance of X (underdispersion); and for $k = 0$, $\mu = VX$, and the GP distribution is reduced to a standard Poisson distribution.

In Chapter 2 we apply a univariate GP model to equidispersed, underdispersed, and overdispersed data sets. This application demonstrates the ability of the GP model to detect equi/under/over dispersion in discrete data. We use Bayesian Methodology to make inference on the parameters (μ, k) for each of these three cases.

In Chapter 3 we consider longitudinal frequency data collected from several individuals. Our application models the n_t frequencies at time t as independent GP observations, with mean defined by a piecewise-linear function of time. Attractive features of this piecewise-linear model include random knot locations as well as the ability to incorporate missing observations. We term this model a piecewise linear GP

regression model. We use the results from this model to make comparisons between experimental groups. A discussion of these results, as well as of the GP distribution and its limitations, is found in Chapter 4.

Chapter 2

Generalized Poisson Application

In this chapter we develop a basic model to make inference on the parameters μ and k of a GP distribution. This inference is based on univariate data assumed to come from a GP distribution. To this end, an underdispersed, an equidispersed, and an overdispersed data set were simulated using the $S-plus$ software package. Making these data sets requires little more than discretizing the unit interval into segments whose lengths are equal to the values of $f(x|\mu,k)$ from (1.2). Then a random uniform (0,1) draw falls into one of these segments and hence defines an (integer) realization from the GP distribution. Each additional random uniform (0,1) draw is used in the same manner to generate an additional GP realization. For more background on simulation from discrete (or continuous) distributions, see Ross 2003. This simulation process yields a data set $y_1, y_2, ..., y_n$ of iid GP random variables that are either underdispersed, equidispersed or overdispersed, depending on the value chosen for the parameter k.

To model frequency data assumed to have come from a GP distribution, we use the (μ, k) parameterization found in the density from (1.2). Thus, the log likelihood function is given by

$$l(\mu, k) = \sum_{j=1}^{n} \left[y_j \log\left(\frac{\mu}{1+k\mu}\right) + (y_j - 1)\log(1 + ky_j) - \log(y_j!) - \frac{\mu(1+ky_j)}{1+k\mu} \right]. \quad (2.1)$$

In the next section we discuss how Bayesian inference on the parameters μ and k is made.

2.1 Univariate Parameter Estimation

In the case of this GP application there are only two parameters that must be estimated, the mean μ and the dispersion parameter k. We employ Markov Chain Monte Carlo (MCMC) techniques using a Metropolis-Hastings update step to estimate each parameter. For more on Metropolis Hastings updating see Gilks, et.al. 1996.

We 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 the current value $\mu^{(0)}$.

$$\mu^* \sim \text{Unif}(\mu^{(0)} - a, \mu^{(0)} + a), \quad (2.2)$$

where $a > 0$ is a tuning parameter in the proposal distribution for μ^*. We have let the proposal distribution for μ^* be uniform (2.2) which gives

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

The minimum function in the denominator gives the length of the proposal interval, which may be constrained by zero or the value of k. In addition, $\frac{q(\mu^{(0)}|\mu^*)}{q(\mu^*|\mu^{(0)})}$ is the Hastings ratio based on the proposal distributions.

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

$$\alpha = \min\left(1, \frac{L(\mu^*, k)\pi(\mu^*, k)q(\mu^{(0)}|\mu^*)}{L(\mu^{(0)}, k)\pi(\mu^{(0)}, k)q(\mu^*|\mu^{(0)})}\right). \tag{2.3}$$

Here L is the (non-logged) likelihood function from (2.1). A (noninformative) joint uniform prior $\pi(\mu^*, k)$ was chosen for all possible (μ, k) combinations; this means the ratio of the prior distributions in (2.3) will equal one.

The following equation gives the logged ratio α in the acceptance probability from (2.3). The logged version is used for purposes of computational efficiency:

$$\begin{aligned}
&\log\frac{L(\mu^*, k)\pi(\mu^*)q(\mu^{(0)}|\mu^*)}{L(\mu^{(0)}, k)\pi(\mu^{(0)})q(\mu^*|\mu^{(0)})} = \\
&\sum_{j=1}^{n}\left[y_j\log\left(\frac{\mu^*}{1+k\mu^*}\right) - \frac{\mu^*(1+ky_j)}{1+k\mu^*} - y_j\log\left(\frac{\mu^{(0)}}{1+k\mu^{(0)}}\right) + \frac{\mu^{(0)}(1+ky_j)}{1+k\mu^{(0)}}\right] \\
&+\log(q(\mu^{(0)}|\mu^*)) - \log(q(\mu^*|\mu^{(0)})).
\end{aligned} \tag{2.4}$$

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

analogous fashion, with the following exceptions:

- The proposed value k^* is drawn from a uniform distribution centered at the current parameter value $k^{(0)}$; the value of the tuning parameter a is allowed to vary from that used in proposing μ^*. Considering all possible constraints on k gives the proposal distribution q as:

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

- The acceptance probability α is then calculated from (2.3). Note, the ratio of the prior distributions equals one just as it did when updating μ because the joint uniform prior was chosen for all (μ, k) combinations. The ratio of the likelihoods, however, is slightly different, as terms involving k that were proportionality constants when updating μ are no longer ignorable. The resulting logged acceptance ratio α is then given by

$$\begin{aligned}\log\frac{L(\mu,k^*)\pi(k^*)q(k^{(0)}|k^*)}{L(\mu,k^{(0)})\pi(k^{(0)})q(k^*|k^{(0)})} = \\ \sum_{j=1}^{n}\left[y_j\log\left(\frac{\mu}{1+k^*\mu}\right)+(y_j-1)(1+k^*y_j)-\frac{\mu(1+k^*y_j)}{1+k^*\mu}-y_j\log\left(\frac{\mu}{1+k^{(0)}\mu}\right)\right. \\ \left.-(y_j-1)(1+k^{(0)}y_j)+\frac{\mu(1+k^{(0)}y_j)}{1+k^{(0)}\mu}\right]+\log(q(k^{(0)}|k^*))-\log(q(k^*|k^{(0)})).\end{aligned}$$

The values $\mu^{(1)}$ and $k^{(1)}$ represent the updated values of these parameters (note that $\mu^{(1)}$ may equal $\mu^{(0)}$ if μ^* was not accepted). These values are then treated as the new "current" parameter values, and the entire updating process is repeated, making sure to store the updated values of μ and k at each iteration. This process is continued until the values stored for $\{\mu, k\}$ have converged to the posterior distribution for $\{\mu, k\}$.

The posterior distribution is *the* inference-making tool, as it is the distribution of the parameters conditional upon the observed data.

The parameters μ and k were updated using a program written in $S-plus$. The program saved the current parameter every 50 iterations to avoid autocorrelation. We now provide the results obtained from applying this model to the three simulated data sets.

2.2 Equidispersion

Here the model is tested on data generated from a GP distribution whose mean is equal to its variance. For this simulation we set the GP mean and variance equal to seven. Extra precaution had to be made when updating the parameters, because in this case the dispersion parameter k would be close to zero. In other words, the value of k changes from positive to negative or vice versa. The computer simulated equidispersed data set of $n = 200$ realizations has mean $\bar{x} = 7.03$ and variance $s^2 = 7.074472$. Letting $\hat{k}$ denote the empirical value of the dispersion parameter based on our sample, we have $\hat{k} = 0.0004492$. Figure 2.1 shows the distribution of this simulated data set.

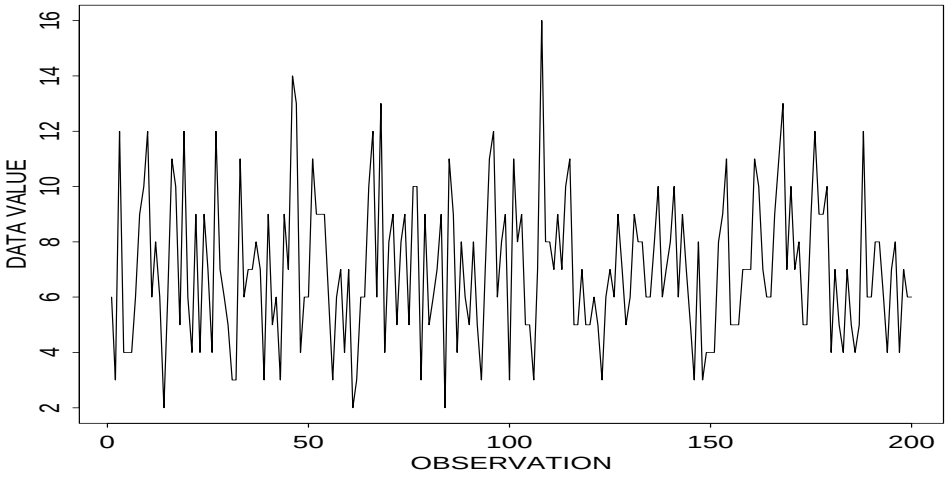

Figure 2.1: Distribution of the computer simulated equidispersed data set.

Figures 2.2 and 2.3 show the posterior trace plots for both parameters. The parameter μ in Figure 2.2 ranges from about 6.5 to 7.5, with a mean of approximately 7. The parameter k, although it falls both above and below zero, still maintains very close to zero. Notice from Figure 2.3 that $P(k \geq 0) = .5$ demonstrating equidispersion.

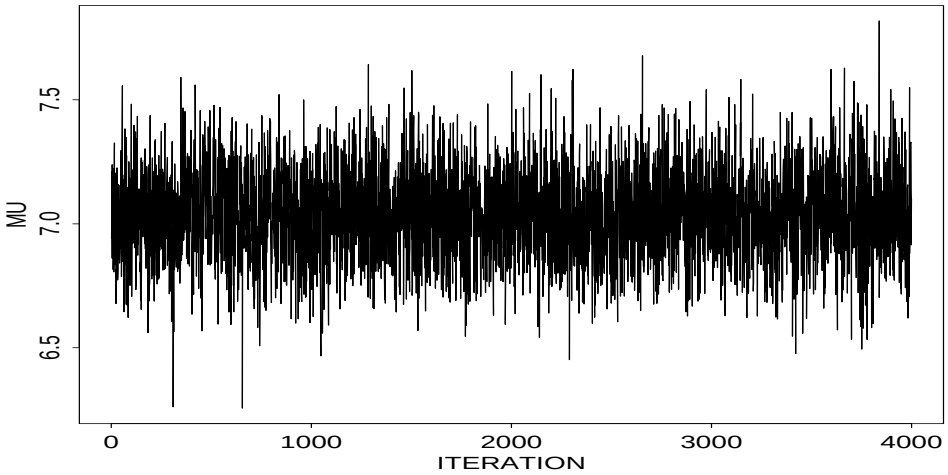

Figure 2.2: Trace plot of the marginal posterior draws for the parameter μ in the equidispersed data set.

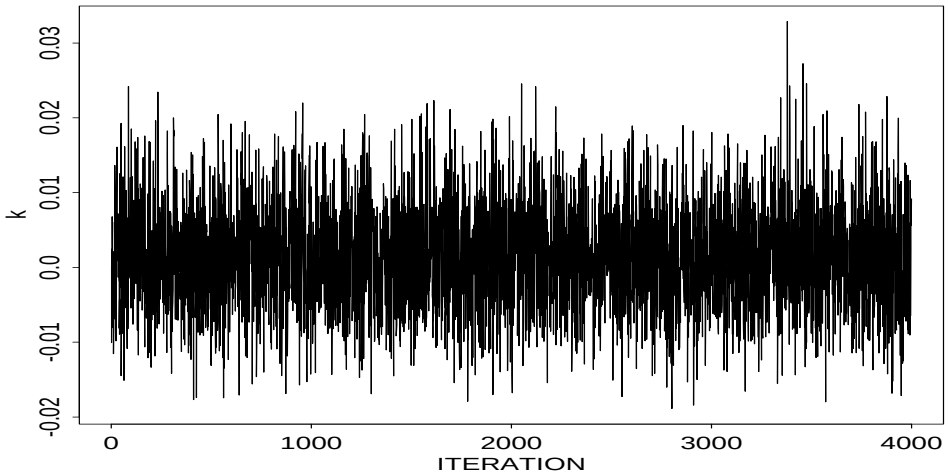

Figure 2.3: Trace plot of the marginal posterior draws for the parameter k in the equidispersed data set.

2.3 Underdispersion

Here the model is applied to the underdispersed simulated data set. Before we begin, a notable fact about underdispersed GP distributions is that the value chosen for k, determines the range of values that μ can take. That range is $0 < \mu < \frac{1}{-2k}$, which can be shown in the following manner:

$$\begin{aligned} k &= \frac{\lambda}{\theta}, && (2.5) \\ \mu &= \frac{\theta}{1-\lambda}, \quad \text{and} \\ -1 &< \lambda < 0, && (2.6) \end{aligned}$$

where λ and θ are the GP parameters given in (1.1). Clearly, from (2.5), $\lambda = k\theta$. By substitution, the value of $k\theta$ must lie between -1 and 0 (2.6). Since k is a negative number in the underdispersion case we arrive at the following inequality: $0 < \theta < -k$. Evaluating $\mu = \frac{\theta}{1-k\theta}$ over the interval for θ gives

$$0 < \mu < \frac{1}{-2k}. \tag{2.7}$$

It is important to note that because the range for μ depends on k, the range for k must depend on μ as well. In the GP model section we determined that the range for k is $\frac{-2}{\mu} < k < 0$, which comes directly from the fact that the variance is less than the mean in underdispersion, or $\mu(1+k\mu)^2 < \mu$. This statement, although true, does not give the complete story. When we determined the above range for μ, we also found that k must abide by the same inequality. In other words,

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

Therefore, there is a limit to how much smaller the variance is allowed to be than the mean. Since $\frac{-1}{2\mu}$ is always closer to zero than $\frac{-2}{\mu}$, we then say that given μ, the value of k is constrained in the interval $\frac{-1}{2\mu} < k < 0$. Further discussion on this topic can be found in Section 4.1.

The following $S-plus$ function was written to show the possible values of μ when the above system of equations is held. The function accepts a dispersion parameter k and the number of iterations. It draws a random number for θ that fits within the possible range of values for θ. Then it uses that number θ to find both λ and μ. In the following code λ_1 is equivalent to θ and λ_2 is equivalent to λ.

```
under <- function(k,n) {
  x.s <- NULL
  for(i in 1:n){
    range <-(1/-k)
    lam1 <- runif(1,0,range)
    lam2 <- lam1*k
    mu <- lam1/(1-lam2)
    x.s <- c(x.s,mu)
  }
  return(x.s)
}
```

The possible μ values obtained from the above function, for $k = -0.04$ and $10,000$ iterations, are displayed in Figure 2.4. Note from (2.7) that $0 < \mu < 12.5$ is the range

under this particular value of k; this is supported by the simulation.

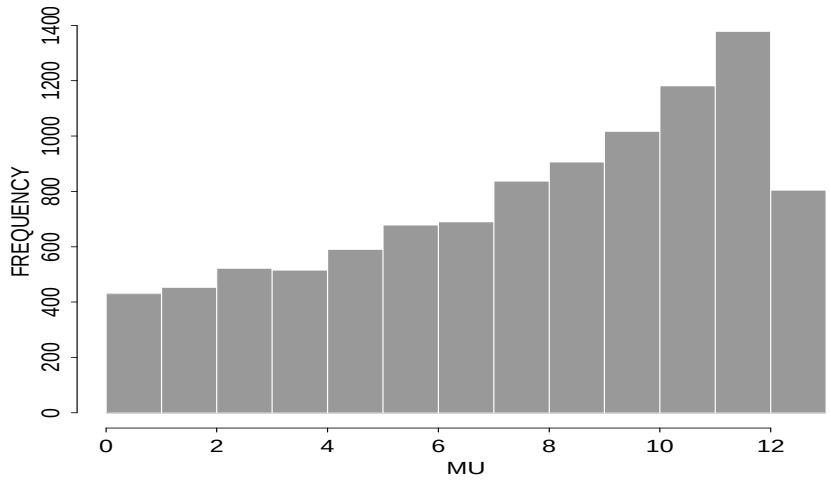

Figure 2.4: Histogram of the range of possible values of μ when $k = -0.04$.

The computer simulated underdispersed data set of $n = 200$ realizations was generated from a GP distribution with mean $\mu = 8$ and dispersion parameter $k = -0.04$. Figure 2.5 shows the distribution of this data set with mean $\bar{x} = 7.84$, variance $s^2 = 3.622513$, and dispersion parameter $\hat{k} = -0.0408486677$.

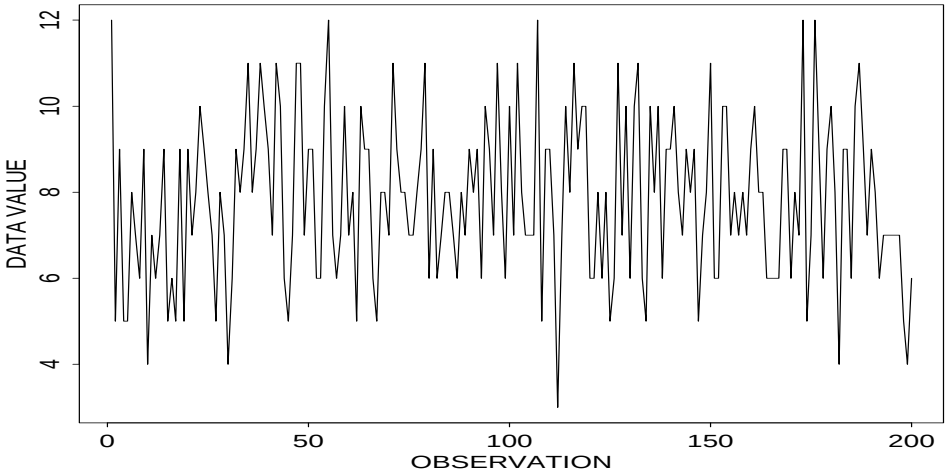

Figure 2.5: Distribution of the computer simulated underdispersed data set.

Figures 2.6 and 2.7 show the posterior trace plots for both parameters based on the computer simulated underdispersed data set. The parameter μ falls mainly between

the values of 7.5 and 8.2, which is close to the mean of the actual sample. The parameter k also remains close to the value of $\hat{k}$ and lies mainly between -0.05 and -0.03. The $P(k \geq 0) = 0$ demonstrating underdispersion; this is supported by Figure 2.7.

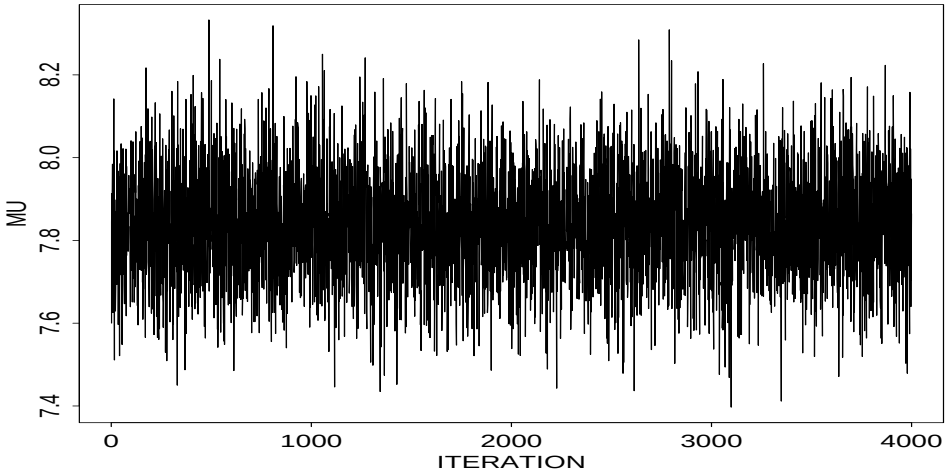

Figure 2.6: Trace plot of the marginal posterior draws for μ in the underdispersed data set.

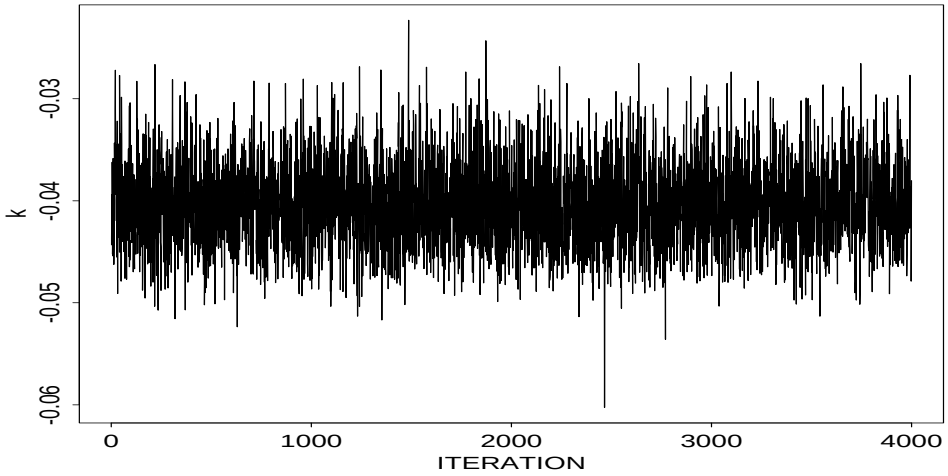

Figure 2.7: Trace plot of the marginal posterior draws for k in the underdispersed data set.

2.4 Overdispersion

The model is now tested on data generated from a GP distribution whose variance is greater than its mean. For this simulation we set the mean $\mu = 10$ and dispersion parameter $k = 0.03$. The computer simulated underdispersed data set of $n = 200$ realizations has mean $\bar{x} = 10.05$, variance $s^2 = 17.50503$, and dispersion parameter $\hat{k} = 0.03181794$. Figure 2.8 shows the distribution of this data set.

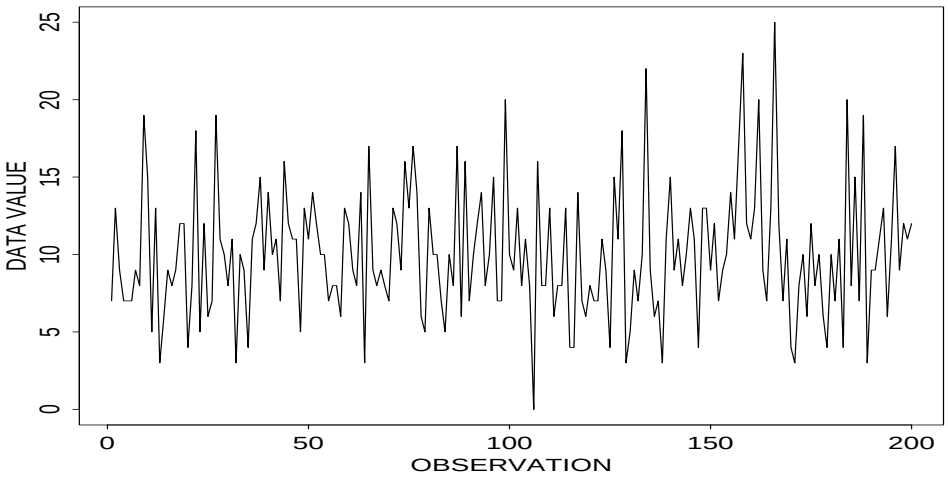

Figure 2.8: Distribution of the computer simulated overdispersed data set.

Figures 2.9 and 2.10 show the posterior trace plots for both parameters based on the computer simulated overdispersed data set. The parameter μ falls mainly between the values of 9.5 and 10.5, which is close to the mean of the actual sample. The parameter k also remains close to the the actual value of the sample data. The $P(k \geq 0) = 1$ demonstrating overdispersion; this is supported by Figure 2.10.

In this chapter we showed the GP distributions ability to detect equi/under/overdispersion. In the next chapter we define our piecewise-linear GP model and apply it to real data collected from menopausal women to make comparisons across experimental groups.

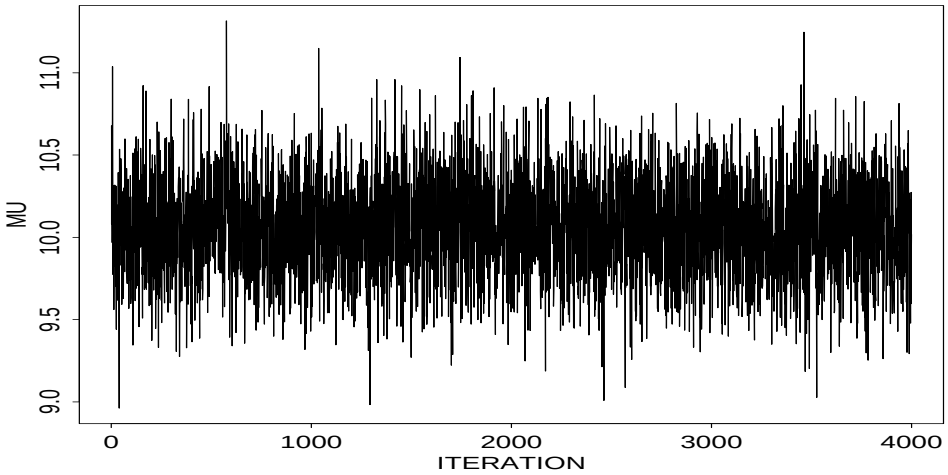

Figure 2.9: Trace plot of the marginal posterior draws for the parameter μ in the overdispersed data set.

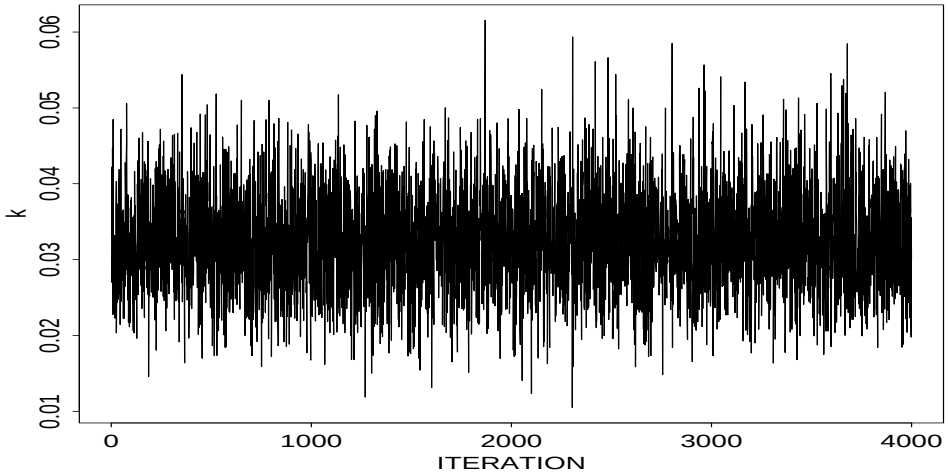

Figure 2.10: Trace plot of the marginal posterior draws for the parameter k in the overdispersed data set.

Chapter 3

Hot Flush Application

3.1 Data

The data used in this application comes from a clinical trial investigating acupuncture as an alternative treatment to alleviate symptoms of menopause for breast cancer survivors. Due to the risk of recurrence, traditional hormone therapy is an unfavorable option for these cancer surviving menopausal women. The women were split into three groups: a control group ($n = 17$), a treatment group ($n = 16$), and an educational group ($n = 6$). The women in the treatment group were given acupuncture in effective areas, while the women in the control group were given acupuncture in supposedly ineffective areas. The women in the educational group were enrolled in a weekly education class that explained menopause effects and offered advice on healthy living. Women from each group reported the number of hot flushes they experienced per day for a total of 91 days. This time period includes an initial baseline week, during which no treatment or education was administered. Figure 3.1 shows the profile of one of the individuals in the treatment group.

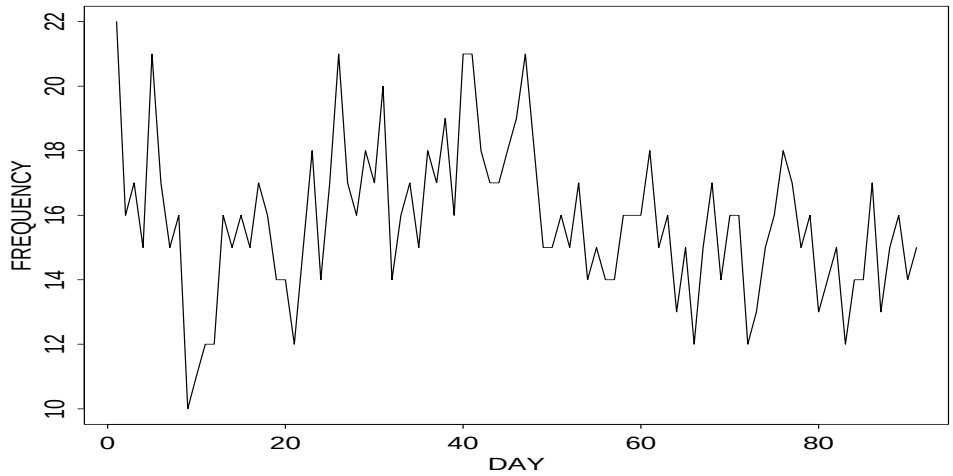

Figure 3.1: Daily hot flush frequencies experienced by a subject in the treatment group.

3.2 Data Model

In the spirit of Kern and Cohen 2003, we model longitudinal frequency responses from multiple individuals using a GP distribution whose mean μ is a function of time. We denote the mean of the GP distribution at time t as μ_t. Thus, if the n_t frequencies $\{y_1, \ldots, y_{n_t}\}$ recorded at time t are modeled as GP with mean μ_t and dispersion parameter k, then from (2.1) the log-likelihood function $l_t(\mu_t, k)$ for these n_t observations is:

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

The log-likelihood for the observations at two time periods $t = 1$ and $t = 2$ is

$$l_1(\mu_1, k) + l_2(\mu_2, k),$$

and thus the log-likelihood $L(\boldsymbol{\mu}, k)$ for all time points $t = 1, 2, ..., 91$ is

$$l(\boldsymbol{\mu}, k) = \sum_{t=1}^{91} l_t(\mu_t, k),$$

where $\boldsymbol{\mu} = \{\mu_1, \mu_2, ..., \mu_{91}\}$.

The following section details the piecewise-linear function used to model the relationship between frequency and time.

3.3 Piecewise Linear Function

Flexibility in allowing μ_t to vary with time is obtained by modeling μ_t as a piecewise linear function of time (rather than restricting μ_t to a functional form such as e^{-t}). Specifically, we define a vector of knot locations $\boldsymbol{K} = \{K_1, K_2, \ldots, K_m\}$ and a vector of corresponding heights $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ such that the coordinate (K_i, λ_i) represents the location at which two line segments meet. The (K_i, λ_i) coordinate on the time-frequency plane is referred to as a *node*. This yields the following:

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

for $i = 1, ..., m$. The above function is simply the point slope form of a line with slope $\left(\frac{\lambda_{i+1}-\lambda_i}{K_{i+1}-Ki}\right)$ passing through the points (K_i, λ_i) and (K_{i+1}, λ_{i+1}), where K_i is a specific knot location on the horizontal (time) axis and λ_i is a specific height on the vertical axis. It is important to note that t must be a time between the knots K_i and K_{i+1}.

Five nodes were selected to formulate the piecewise linear function ($m = 5$): three at fixed locations and two at random knot locations. Intuitive choices of the fixed knot locations were $\{0.5, 7.5, 91.5\}$, where the baseline week is simply separated from the remaining weeks of treatment. These times correspond to the beginning of the

study, the end of the baseline week, and the end of the study respectively. The two additional knot locations were randomly chosen in the time interval $(7.5, 91.5)$ to give additional flexibility to the piecewise linear model. Two random locations avoid over parameterizing the model, but still allow more than one line segment to represent each of the μ_i's within the time interval $(7.5, 91.5)$. Utilizing the same notation from above gives the vector of knot locations $\boldsymbol{K} = \{K_1, K_2, \ldots, K_5\}$ and vector of node heights $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_5\}$, where $K_1 < K_2 < K_3 < K_4 < K_5$ and $\lambda_i > 0$. Note that K_1, K_2 and K_5 are the fixed knot locations; K_3 and K_4 are modeled as random. These eight parameters (the two random knot locations, the five node heights, and the dispersion parameter k), completely describe our piecewise linear GP regression model. Further detail will now be given on how inference on each of these parameters is made.

3.4 Multivariate Parameter Estimation

Let

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

represent the "current" value of the parameters. Note $i = 0$ represents the initial parameter values, where i is simply an index. The parameters were updated as follows, using the Metropolis Hastings algorithm described in Section 2.1.

- To update λ_1 we use a manner similar to updating μ and k in the univariate case. We let the proposal distribution be uniform over an interval centered around the current value $\lambda_1^{(i)}$. Let λ_1^* represent the proposed value of $\lambda_1^{(i)}$. 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

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

Here L is the (non-logged) likelihood function from (3.1), where $\boldsymbol{\mu}$ is expressed in terms of $\boldsymbol{K}$ and $\lambda_1, \ldots, \lambda_5$. The accepted parameter value (λ_1^* with probability α, $\lambda_1^{(i)}$ with probability $1 - \alpha$) is treated as a new "current" value. The other seven parameters are then updated using the "current" value of λ_1.

- The remaining node heights, $\lambda_2, \ldots, \lambda_5$, 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^* is constrained to 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 (i.e. another parameter value is proposed from q, and if accepted—with probability α—stored as $\lambda_1^{(i+1)}$). This process is continued until the values stored for $\{\boldsymbol{\lambda}^{(i+1)}, \boldsymbol{K}^{(i+1)}, k^{(i+1)}\}$ have converged to the posterior distribution for $\{\boldsymbol{\lambda}, \boldsymbol{K}, k\}$.

These parameters were repeatedly updated using a customized program in C. Current parameter values were saved every 250 iterations to avoid auto correlation.

3.5 Results

Figures 3.2, 3.3, and 3.4, show trace plots of the dispersion parameter k for the treatment, placebo, and education group respectively. The probability that the dispersion parameter k is greater than zero is 1 for all three cases, thus demonstrating overdispersion in each group. This is supported by the trace plot for each experimental group. Figure 3.2 shows the posterior mean for k is approximately 0.28 for the treatment group. The posterior mean for k is approximately 0.16 for the placebo group (see Figure 3.3) and 0.10 for the education group (Figure 3.4). All of these figures provide strong evidence suggesting overdispersion in the daily hot flush frequencies.

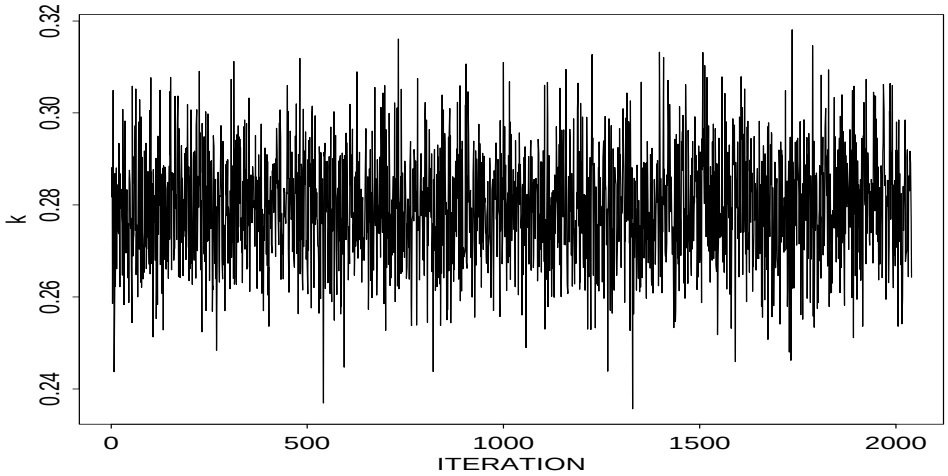

Figure 3.2: Plot of the marginal posterior draws for the parameter k based on analysis of the treatment group data.

Figures 3.5, 3.6, and 3.7 display the results of the piecewise linear GP model individually applied to the treatment, placebo, and education data respectively. From these three graphs we can see the differences across groups. Expected average of hot flush frequency drops by almost three for the treatment group, whereas the placebo group drops by a little over two. There seems to be no noticeable increase or decrease in the expected average hot flush frequency for the education group. Also note that the confidence limits are more narrow for the treatment and placebo group than they are

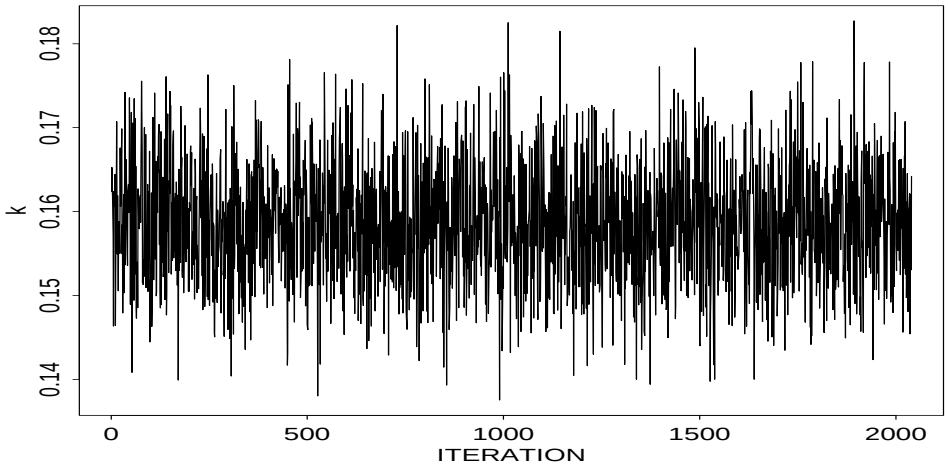

Figure 3.3: Plot of the marginal posterior draws for the parameter k based on analysis of the placebo group data.

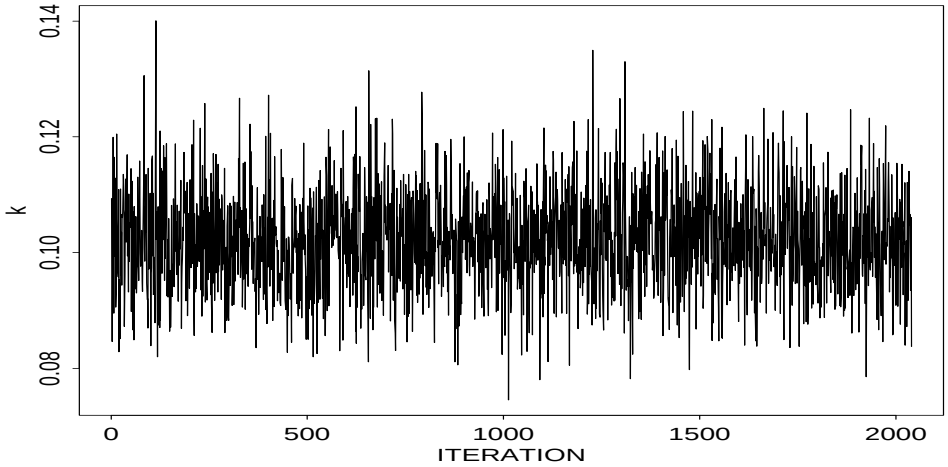

Figure 3.4: Plot of the marginal posterior draws for the parameter k based on analysis of the education group data.

for the education group. This is due to the small number of subjects in the education group (6) relative to the other groups (16 and 17 in the treatment and placebo group respectively).

According to Figures 3.5 and 3.6, there is little difference in the drop of mean hot flush frequencies between the treatment and placebo group. In fact, comparing these two groups to the education group (Figure 3.7) suggests the possibility of a placebo effect. The mean hot flush frequency drops over the period of 91 days for women who

received acupuncture. It is seemingly unimportant as to whether that acupuncture was received in effective areas as opposed to ineffective areas.

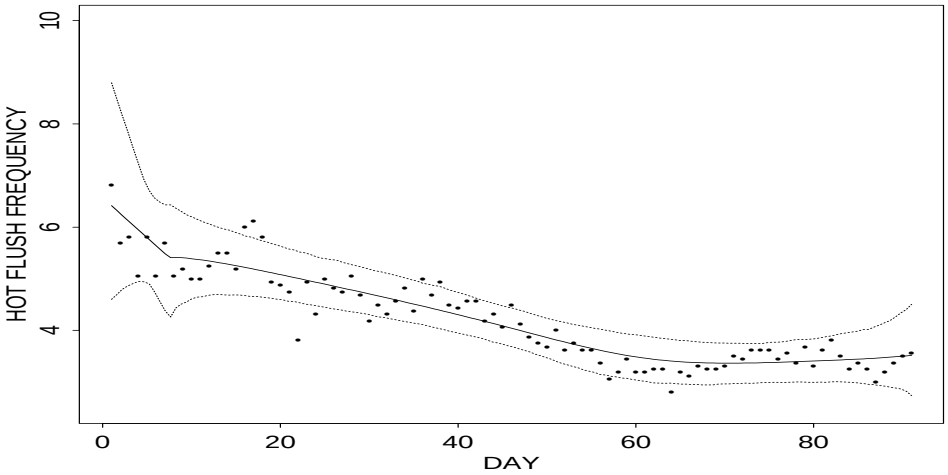

Figure 3.5: Posterior mean hot flush frequency (solid line) with upper and lower confidence limits (dashed lines) for the treatment group. Scatterpoints represent actual daily HFF means.

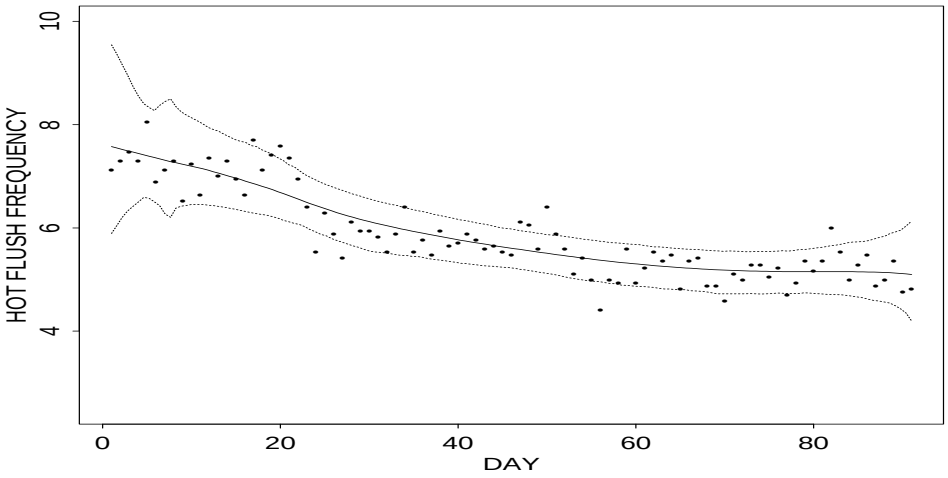

Figure 3.6: Predicted mean hot flush frequency (solid line) with upper and lower confidence limits (dashed lines) for the placebo group. Scatterpoints represent actual daily HFF means.

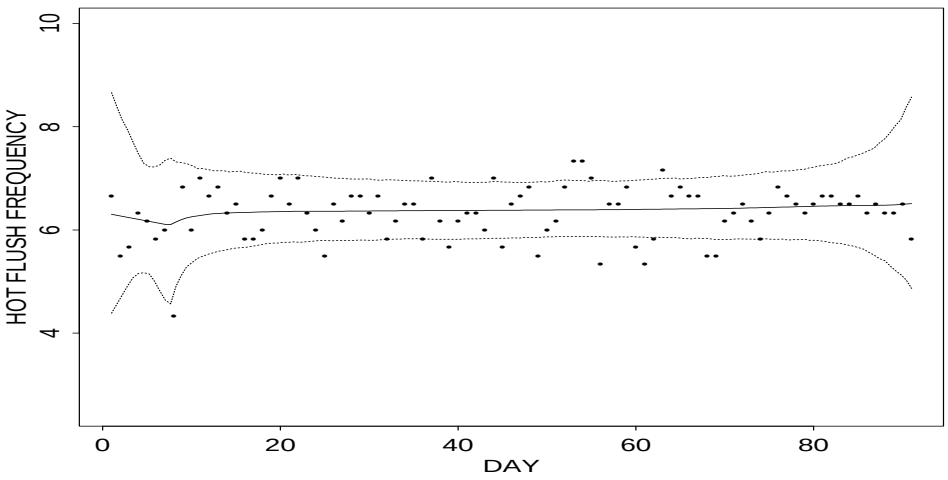

Figure 3.7: Predicted mean hot flush frequency (solid line) with upper and lower confidence limits (dashed line) for the education group. Scatterpoints represent actual daily HFF means.

Chapter 4

Discussion

In this analysis we have constructed a GP model for univariate frequency data and demonstrated the usefulness of this model in allowing for and detecting the dispersion characteristics of the data. We then applied a piecewise linear GP regression model to longitudinal frequency data self-reported from multiple individuals to make comparisons across experimental groups.

Useful features of this piecewise-linear model include its ability to recognize the discrete nature of the data, and adapt to equi/under/overdispersion exhibited in the data. It also allows for time correlation of the daily means through the piecewise linear function for the mean.

This model is useful in other applications where frequency data is collected across time and individuals. Application of this model on a different data set would require specification of the number of nodes, including the number of random and fixed knot locations. In our application, we decided upon five nodes, three fixed and two random knot locations, to provide flexibility without over parameterizing the model. It is important to note that the heights (λ_i's) have the potential to be fixed as well. Consider, for example, setting the height at the start of the baseline week equal to

the height at the end of the baseline week ($\lambda_1 = \lambda_2$) in the hot flush frequency study. This is appropriate if one assumes that the average hot flush frequency is constant over the baseline week. We decided against setting $\lambda_1 = \lambda_2$ because of potential self-monitoring effects.

It must be noted, however, that this model treats data from two separate time points as independent when in our application the data at two separate time points comes from the same individuals. Further study using a different model that explicitly incorporates the dependence structure of the data could be implemented and the results obtained compared to the results of this model.

4.1 Limitations to Underdispersion

Although the GP distribution models both underdispersion and overdispersion, there are limitations to how underdispersed a GP random variable can be. As an example, consider the highly underdispersed frequency data from one of the subjects in the placebo data set. Figure 4.1 shows the distribution of the patient's hot flush frequencies over the 91 days. The mean of the subject's hot flush frequencies is 13.73626 and the variance is 1.640781.

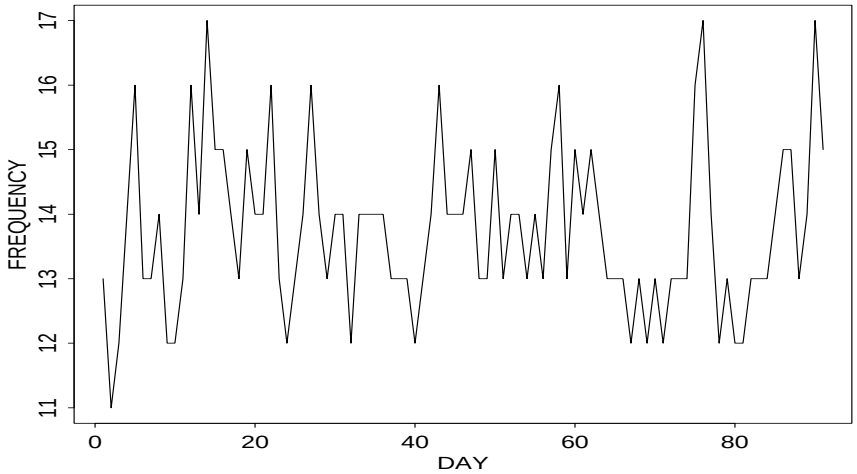

Figure 4.1: Daily hot flush frequencies experienced by a subject in the placebo group.

One might assume that these 91 frequencies come from the GP distribution. However, note that in this case the empirical value of the dispersion parameter is $\hat{k} = -0.0478$. The value found for k follows directly from 1.3 by substituting for μ and VX the sample mean and the sample variance of this subject's frequency data. If the dispersion parameter k is set to -0.0478, the upper bound for μ from (2.7) is the following:

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

Yet not only does the mean of the data ($\bar{x} = 13.73626$) exceed this upper bound, every observation in this data set is greater than the upper bound. We therefore can quite conclusively determine that the data could not have come from the GP distribution.

4.2 Future Work

Future work using a GP model for discrete data includes specifying a non-uniform prior distribution on k. For example, let k have a mixture distribution that places positive probability on zero and on the real numbers. In this way the posterior probability that $k = 0$ can be directly estimated (giving the probability of equidispersion). In similar fashion the posterior probability that $k > 0$ or $k < 0$ would provide estimates of the chances of over/underdispersion, respectively.

Additional further study would be to compare our piecewise GP regression model with alternative models for longitudinal frequency data. For example, implementing a model that explicitly incorporates the dependence structure of the data and comparing the results with those from our model would provide more insight to successful longitudinal data modeling strategies.

References

Consul, P.C. and G.C. Jain. 1973. A generalization of the Poisson distribution. *Technometrics* 15: 791-799.

Famoye, F. 1993. Restricted generalized Poisson regression model. *Commun Statist.— Theory and Meth.* 22(5): 1335-1354.

Famoye, F. and W. Wang. 1997. Modeling household fertility decisions with the generalized Poisson regression. *Journal of Population Economics* 10: 273-283.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice — Interdisciplinary Statistics* London: Chapman & Hall.

Kern, J. and S.M. Cohen. 2003. Menopausal symptom relief with acupuncture: modeling longitudinal frequency data. submitted to: *Bayesian Analysis.*

Ross S. 2002. *Simulation 3rd Edition* San Diego, California: Academic Press.

Weisstein, E.W. 1999. *CRC Concise Encyclopedia of Mathematics.* Boca Raton: Chapman and Hall.