ABSTRACT

Spatial Poisson Regression:
Bayesian Approach Correcting for Measurement Error with Applications

William H. Atkinson, Ph.D.

Chairperson: Thomas L. Bratcher, Ph.D.

Under and over reporting is a common problem in social science research, adverse events associated with drug use, and many other areas of research. Furthermore, overdispersion is another common problem that plagues count data. McBride (2006) proposed a Bayesian Poisson regression model which accounts for overdispersion in count data. We extend this model by adding parameters to accommodate the problems associated with under and over reporting in the count data. We then study the model's coverage, power, accuracy of point estimates, and credible set widths through simulation using a spatial lattice grid. We find that our proposed model produces reliable point estimates and reasonable credible set widths, coverage, and power.

We also provide two examples of the models use: disease mapping of habitat burglary from the city of Waco Texas and an analysis of sports data similar to that of Albert's (1992) analysis of homerun data. Research questions of interest are answered using the subset selection procedure proposed by Bratcher and Bhalla (1974), used by Hamilton, Bratcher, and Stamey (2008) and Stamey, Bratcher, and Young (2004), to demonstrate the ease of use for combining the our model developed here and the subset selection procedure itself, as was also done in McBride (2006).

Spatial Poisson Regression:
Bayesian Approach Correcting for Measurement Error with Applications

by

William H. Atkinson, B.S., M.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

Thomas L. Bratcher, Ph.D., Chairperson

_____

John W. Seaman, Ph.D.

_____

James Stamey, Ph.D.

_____

Jack D. Tubbs, Ph.D.

_____

Steve Green, Ph.D.

Accepted by the Graduate School
August 2010

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

To Greg Miller, Ph.D., William Clark, Ph.D, Wayne Proctor, Ph.D., and the late Jasper Adams, Ph.D., I say thank you for pushing me so hard. I'd also like to thank Tom Carlile for all the help in the early stages of my journey. To Tom Bratcher, Ph.D., I say thank you for allowing me to take on this project independently and giving me the freedom and support to see it through. Thank you for your tireless editing and helpful comments.

DEDICATION

To our Founding Fathers for the blessings of liberty we enjoy in our republic.

To the Freedom we have to better ourselves and our lives.

To those that believed in me even when I would not: mother, father, sister, Roger, and most of all my Lord and Savior.

CHAPTER ONE

Introduction

As researchers we are interested in the behavior of a system under study and often wish to establish a *cause and effect* relationship between some response variables and some treatment variable of interest. This process leads to experiments which produce data. From these data, assuming ideal conditions, we are able to answer certain questions of interests concerning the system under study. However, *ideal* conditions are not always available due to the nature of problem. Consider for instance trying to establish a causal effect of drinking paint and death. Clearly ethical issues will arise in this otherwise pointless study.

In such situations we as researchers are left to rely on observational data of which we are only able to answer limited questions and speculate as to the causal relationship between response and treatment. Furthermore, observational data are often riddled with errors of various sources: recall bias, over/under reporting, and differing methods of collection over time, often the case with government data, just to name a few. Observational data are often recorded in the form of event counts, a non-negative integer valued observation describing the number of times a particular event occurred during a specified time period or within a certain region. For example, every ten years the United States Census collects count data of several responses on the United States citizenry broken down by Census tracts.

Many applications of public health data, phase IV safety trials, criminology, econometrics, and ecology studies all depend on data of a count nature. For example, the number of adverse events attributed to a particular drug reported to regulatory bodies and the number of traffic accidents at a certain intersection just to name two. Both of these examples are of a count nature and the complexity of the issue stems

from the research question of interest, typically attempting to associate the counts with various predictor variables.

This process of associating a response variable with potential predictor variables, often termed *regression analysis*, is not new and can be dated to Sir Francis Galton (1886). Galton studied the relationship between the heights of parents and children and commented that the heights of children, from both tall and short parents, seemed to regress to the mean of their respective group.

Analysis of count data using discrete parametric distributions for univariate random variables has a rich history and many applications (Johnson, Kotz, and Kemp, 1992). One very often used distribution for count data analysis is the Poisson distribution which was originally developed as the limiting case of the binomial (Poisson, 1837). A historically well known application using the Poisson distribution is the analysis performed by Bortkiewicz (1898), a study of the annual deaths of Prussian soldiers due to being kicked by mules. Another example is Albert's (1992) paper where he modeled the homerun hitting rates of historically famous baseball players such as Micky Mantle and Babe Ruth. Spanning now over a century of application and research we need only to search the literature and standard texts briefly for countless uses of the Poisson distribution.

Count and event-location data have been used extensively in epidemiology and disease mapping research. A classic example in which maps and the idea of clustering was used goes back to the 1854 London cholera epidemic. Snow (1854) hypothesized cholera was spread through contaminated drinking water after having extensively mapped the cases. Plam (1890) showed that rickets occurred in locations with cold, wet climates. Blum (1948) study where he used mapping to show that sunlight is a factor in the onset of skin cancer. Today, more sophisticated modeling and statistical methods specifically developed for spatial data are available to researchers tackling complex problems.

Volumes of work have been written on the subject of regression analysis for count data, spatial statistics, and many texts have at least a section discussing basic applications and refers the reader elsewhere. Here we refer the reader to the following works for more information (Schabenberger and Gotway 2005; Cressie 1993; Banerjee, Carlin, and Gelfand 2004; Lawson 2009; Cameron and Trivedi 1998).

In this dissertation we modify and expand the Bayesian hierarchical Poisson regression model developed by McBride (2006) to include misclassification and measurement error, study some of the model's properties through simulations, and provide two examples of it's use: disease mapping of habitat burglary from the city of Waco Texas and sports data analysis similar to that of Albert's (1992) analysis of homerun data. We then combine the the model with an optimal subset selection procedure developed by Bratcher and Bhalla (1974), used by Hamilton et al. (2008) and Stamey et al. (2004), to answer various research questions of interest.

1.1 Count Regression Models

The Poisson distribution is typically the benchmark univariate distribution used for the analysis of count data. A consequence of assuming a Poisson distribution is the equality of mean and variance. However, this consequence is quite restrictive and is often violated because of *overdispersion* (*under dispersion*) where the variance of the data is greater (smaller) than the mean. One way to correct for this phenomenon is to model the mean of the Poisson distribution, otherwise known as the *rate of occurrence* or *intensity rate* (*rate* for short), as a function of covariates, thus allowing for the mean to change. As a result the conditional variance will change as well. However, conditional on the covariates the Poisson distribution still requires equality of the mean and variance. This assumption, even after the inclusion of covariates, is still often violated.

This problem has many statistical researchers developing models to allow for a relaxation of this assumption. What we provide here in this section is a brief review of several proposed count regression models that are often used. This is not a comprehensive inclusion of all such models but a good basis and starting point for understanding how to correct for *over/under* dispersion. Cameron and Trivedi (1998) discusses these models, as well as others, and also provides references to other sources of interest concerning model properties and applications.

1.1.1 The Benchmark

Assume that for the i^{th} sample or situation the response Y_i is distributed as a Poisson random variable with probability mass function,

$$f(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, y_i \in \mathbb{Z}^+, \tag{1.1}$$

where the rate parameter $\lambda_i = E(Y_i)$. The parameter λ_i may be modeled using covariates but this requires the choice of a link function relating the mean, λ_i, to the chosen covariates. A commonly used link function is the conditional exponential mean function known as the *log-link*. That is, $\lambda_i = exp(\mathbf{x}_i\boldsymbol{\beta})$, or,

$$log(\lambda_i) = \mathbf{x}_i\boldsymbol{\beta}. \tag{1.2}$$

The log-likelihood, under the assumption of independent observations, is given as,

$$ln(L(\boldsymbol{\beta}|\mathbf{x})) = \sum_{i=1}^{n} \left(y_i\mathbf{x}_i'\boldsymbol{\beta} - exp(\mathbf{x}_i'\boldsymbol{\beta}) - ln(y_i!)\right). \tag{1.3}$$

This leads to the first-order conditions,

$$\sum_{i=1}^{n} \left(y_i - exp(\mathbf{x}_i'\boldsymbol{\beta})\right)\mathbf{x}_i = 0, \tag{1.4}$$

which the maximum likelihood estimator for $\boldsymbol{\beta}$ is the solution, $\widehat{\boldsymbol{\beta}}$, to (1.4). Often the solution is found using the Newton-Raphson iterative method. Convergence is guaranteed since the log-likelihood is globally concave. If the actual data generating process is indeed Poisson, we may appeal to asymptotic results and the usual maximum likelihood theory resulting in, $\widehat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where,

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} (\lambda_i \mathbf{x}_i \mathbf{x}_i') \right)^{-1}. \tag{1.5}$$

Statistical inference can then follow and research questions of interest concerning the system under investigation can be answered. Valid statistical inference requires correct specification of both conditional mean and conditional variance. Under the Poisson assumption this requires equality of both mean and variance.

It is well known that even if the assumption $Y_i \sim Poisson(\lambda_i)$ is violated, the MLE for $\boldsymbol{\beta}$ is still consistent assuming the the conditional mean is correctly specified (Gourieroux, Monfort, and Trognon, 1984a), resulting from the fact the Poisson distribution is of the exponential family. However, issues arise when statistical inference is attempted. If the data generating process is not Poisson, valid statistical inference is still possible; but methods to correct for the maximum likelihood variance must be used. One such method is to assume that the variance is a scalar multiple of the mean resulting in what Cameron and Trivedi (1986) call the NB-1 model.

1.1.2 The NB-k Models

In the Poisson regression model we must assume equality of variance and mean. This requires that the variance $V(Y_i) = E(Y_i) = exp(\mathbf{x}_i\boldsymbol{\beta})$. Here we relax this assumption and allow the variance to take on a differing form. The NB-1 and NB-2 models are similar with the only difference being the functional form of the conditional variance. The NB-1 model assumes that the variance of Y_i is a scalar multiple of the mean while the NB-2 model assumes that the variance of Y_i is a quadratic function of the mean.

For notational purposes let $\omega_i = V(Y_i|\mathbf{x}_i)$ denote the conditional variance of Y_i. Since we are discussing functions, let $\omega(\lambda_i, \alpha) = \omega_i$. The NB-k model assumes the functional form for ω to be,

$$\omega(\lambda_i, \alpha) = \lambda_i + \alpha\lambda_i^k, \tag{1.6}$$

where the constant k is specified. Notice that for $\alpha = 0$ we obtain the rather restrictive Poisson variance form discussed in section 1.1.1. The NB-1 specifies $k = 1$ while the NB-2 specifies $k = 2$. For the NB-1 model we end up with the functional form of the variance to be, $(1 + \alpha)\lambda_i$, which is simply a constant multiple of the mean. In practice the parameter α must be estimated using the data at hand. For the NB-2 model we end up with the functional form of the variance to be, $\lambda_i(1+\alpha\lambda_i)$, and again α must be estimated.

Using results from Gourieroux, Monfort, and Trognon (1984b), Cameron and Trivedi (1986) provide the asymptotic distribution of the *pseudo*-MLE as, $\widehat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where,

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} \left(\frac{1}{n}\sum (\lambda_i \mathbf{x}_i \mathbf{x}_i')\right)^{-1} \left(\frac{1}{n}\sum (\omega_i \mathbf{x}_i \mathbf{x}_i')\right) \left(\frac{1}{n}\sum (\lambda_i \mathbf{x}_i \mathbf{x}_i')\right)^{-1}. \tag{1.7}$$

From this formulation it's clear that if the real variance, ω_i, is equal to the variance dictated by the assumed Poisson density, λ_i, then (1.7) reverts to (1.5). Thus the usual maximum likelihood inference is valid.

Focusing on the NB-1 model we can consider the functional form of the conditional variance to be, $(1 + \alpha)\lambda_i = \phi\lambda_i$, where $\phi = (1 + \alpha)$. As mentioned earlier, the assumption that the data generating process be Poisson can be relaxed and still maintain consistency (Gourieroux, Monfort, and Trognon, 1984a). Given this robustness we may still use the Poisson density to obtain the first-order conditions arriving at an estimator if we use the Poisson density for a data generating process that is not truly Poisson. This estimator is known as the *pseudo*-MLE or *qusi*-MLE as it's not truly an MLE for the Poisson distribution but is derived from the Poisson density. The resulting variance structure for $\boldsymbol{\Sigma}$ under the NB-1 model is,

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} \left(\frac{1}{n}\sum (\lambda_i \mathbf{x}_i \mathbf{x}_i')\right)^{-1} \phi, \tag{1.8}$$

since $\omega_i = \phi\lambda_i$. The standard estimator for ϕ is,

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{\left(y_i - \widehat{\lambda}_i\right)^2}{\widehat{\lambda}_i}, \tag{1.9}$$

where p is the number of parameters to be estimated. The motivation for this estimator is that the variance function, $\omega(\lambda_i, \alpha) = \phi\lambda_i$, and thus, $\phi = E(\frac{(y_i-\lambda_i)^2}{\lambda_i})$. The division by $n - p$ rather than n is a degrees of freedom correction and the reader is refereed to Cameron and Trivedi (1998) for more information regarding this estimator.

Turning our attention now to the NB-2 model, Cameron and Trivedi (1998) provides the variance for the psuedo MLE as,

$$\mathbf{\Sigma} = \lim_{n\to\infty} \frac{1}{n^3} \left(\sum \left(\lambda_i \mathbf{x}_i \mathbf{x}_i'\right)\right)^{-1} \left(\sum \left((\lambda_i + \alpha\lambda_i^2)\mathbf{x}_i\mathbf{x}_i'\right)\right) \left(\sum \left(\lambda_i \mathbf{x}_i \mathbf{x}_i'\right)\right)^{-1}. \tag{1.10}$$

A method of moments estimator for α is motivated by the fact that, $E((y_i - \lambda_i)^2 - \lambda_i) = \alpha\lambda_i^2$. Solving for α yields, $\alpha = \frac{E((y_i-\lambda_i)^2-\lambda_i)}{\lambda_i^2}$, motivating the estimator,

$$\widehat{\alpha} = \frac{1}{n-p} \sum_{i=1}^{n} \left(\frac{(y_i - \widehat{\lambda}_i^{\,2}) - \widehat{\lambda}_i}{\widehat{\lambda}_i^{\,2}} \right),$$

where p is the number of parameters to be estimated. The estimators provided here for both α and ϕ are not the only estimators proposed. Further estimators for both α, for the NB-2 variance function, and ϕ, for the NB-1 variance function, are available in Cameron and Trivedi (1986).

1.1.3 Unspecified Variance Function

It is possible to have consistent estimators of the variance for the pseudo MLE without specifying the functional form of ω, the variance of the data generating process. One such method is to use the covariance matrix estimator,

$$\mathbf{\Sigma} = \lim_{n\to\infty} \left(\frac{1}{n}\sum \left(\lambda_i \mathbf{x}_i \mathbf{x}_i'\right)\right)^{-1} \left(\frac{1}{n}\sum \left((y_i - \lambda_i)^2 \mathbf{x}_i\mathbf{x}_i'\right)\right) \left(\frac{1}{n}\sum \left(\lambda_i \mathbf{x}_i \mathbf{x}_i'\right)\right)^{-1}.$$

Cameron and Trivedi (1998) discusses this approach briefly and provides references to Eicker (1967), White (1980), and Robinson (1987) for further information concerning estimators of this type.

1.1.4 Poisson GLM

The Poisson density, using the log-link, can be expressed as,

$$f(y_i|\mathbf{x}_i) = exp\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}y_i - exp(\mathbf{x}_i'\boldsymbol{\beta})}{\phi} + c(y_i,\phi)\right), \tag{1.11}$$

where $c(y_i, \phi)$ is a normalizing constant. Using general linear model theory we know that $V(Y_i) = \phi\lambda_i$, which is the variance function of the NB-1 model.

The estimator $\widehat{\boldsymbol{\beta}}$ for the Poisson GLM maximizes the log likelihood corresponding to (1.11), with respect to $\boldsymbol{\beta}$. The first order conditions for this model are,

$$\sum_{i=1}^{n}\left(\frac{1}{\phi}(y_i - exp(\mathbf{x}_i\prime\boldsymbol{\beta}))\mathbf{x}_i\right) = \mathbf{0}, \tag{1.12}$$

which corresponds to (1.4) where ϕ is a weighting. The resulting estimator for the Poisson GLM is the same as the estimator as the Poisson regression model described in section 1.1.1 with the variance matrix,

$$V(\widehat{\boldsymbol{\beta}}) = \left(\sum(\lambda_i\mathbf{x}_i\mathbf{x}_i')\right)^{-1}\phi, \tag{1.13}$$

which corresponds to the NB-1 model. As such, the estimator for ϕ is often chosen to be the consistent estimator found in (1.9). This estimator can be accessed in SAS 9.2 Proc GLM using the *scale = Pearson* command in the model statement.

Another method is to maximize the log-likelihood with respect to both $\boldsymbol{\beta}$ and ϕ. However, this method poses problems due to the nature of the normalizing constant $c(y_i, \phi)$ (Cameron and Trivedi, 1998).

1.2 Measurement Error

Another problem that arises in epidemiology, clinical trials, ecology, and public health research to name a few is what is known as *measurement error* otherwise known as the *error in variables* problem. Considered as early as Berkson (1950) and Cochran (1968) the effect that errors in predictor variables have on estimators has become a major statistical concern. The statistical literature is filled with authors and books written on this specialized subject. Noteworthy works include, but are not limited to, Fuller (1987), Gustafson (2004), Schwartz and Coull (2003), and McGlothlin, Stamey, and Seaman (2008). An example where measurement error may be an issue is when researchers are not able to measure the desired covariate and instead are forced to measure a surrogate variable. This is often the case in sociology when the desired covariate is poverty but is instead measured using infant mortality due to the lack of information or arbitrary definitions of what poverty is based on income (Pridemore, 2008).

Whittemore and Keller (1988) provide a classical example for Berkson measurement error while Richardson and Gilks (1993) and Dellaportas and Stephens (1995) discuss Bayesian approaches for handling both classical and Berkson measurement error in binomial regression. Several books have been written on the topic of measurement error including Fuller (1987), Gustafson (2004), and Carroll (2006).

Roy, Banerjee, and Maiti (2005) extend the research of a misclassified response by including the case in which one or more covariates are also measured imperfectly. They delvelop a maximum likelihood approach to find corrected estimates of the regression coefficients and apply their method using the Life Span Study (LSS), a study consisting of a large cohort of individuals who survived the atomic bombings of Hiroshima and Nagasaki. This data set was also analyzed by Sposto, Preston, Shimizu, and Mabuchi (1992) where the authors attempted to find the effect of radiation dose on cancer mortality correcting for misclassification in deaths.

As researchers, when there is reason to believe that the response and/or predictors could be measured with error, we must then use methods that account for this possibility. Therefore, the model must be adapted appropriately for the type of measurement error suspected. We provide here a discussion of why this is a problem and a brief review of some of the proposed methods for such a task.

1.2.1 Simple Linear Regression

To understand the effect that measurement error poses on the relationship between two variables consider the simple linear regression scenario when the predictor variables X are subject to measurement error. Here we will assume that the response variable Y is related to the covariate X by $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where the errors ϵ_i are distributed as Normal with mean zero and variance σ^2.

We further assume that we are unable to measure X but are able to measure Z where $Z_i = X_i + \delta_i$. We assume that δ is distributed Normal with mean zero and variance σ_z^2. We finally assume that ϵ and δ are independent, X and δ are independent, and that X and ϵ are independent.

Figure 1.1 displays a random generation of y values when $\beta_0 = 3$, $\beta_1 = 3$, and the error terms ϵ and δ are distributed as $N(0, \sigma^2 = 4)$. The graph on the left plots Y vs X and the graph on the right plots Y vs Z. The lines displayed on the graph are the obtained least squares fit and the actual line used to generate the data. Notice the significant bias towards a zero slope in the least squares fit using the surrogate variable Z in lieu of X. As a matter of fact the obtained least squares estimate for β_1 using the real X's is 2.685 as compared to 0.9145 when using the Z's which is a substantial difference in the estimates. We are able to discern the fact that as the measurement error becomes more severe, that is to say the variability of the error increases, the slope becomes severely biased towards zero.

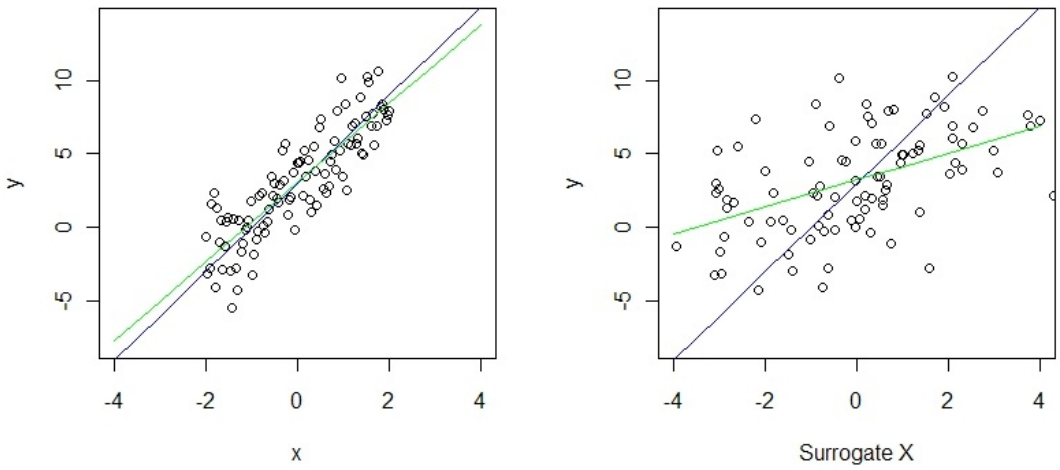

Figure 1.1: Measurement Error: Bias towards Zero.

Another problem is that measurement error often masks features in the data. To illustrate this suppose that $Y = sin(2X) + \epsilon$ with $\epsilon \sim Normal(0, 0.10)$ and $Z = X + \delta$ with $\delta \sim Normal(0, .7)$. Figure 1.2 displays 200 randomly generated points with this structure. As in Figure 1.1 the left graph displays the true X and Y data while the right graph displays the Z and Y data. Notice the sin curve, although clearly visible when the X's are known exactly, is no longer visible in the surrogate data, and there is no clear relationship whatsoever between the response and the surrogate. It is also important to note that in both of these examples the surrogate variable Z is an *unbiased* estimator for the predictor of interest X.

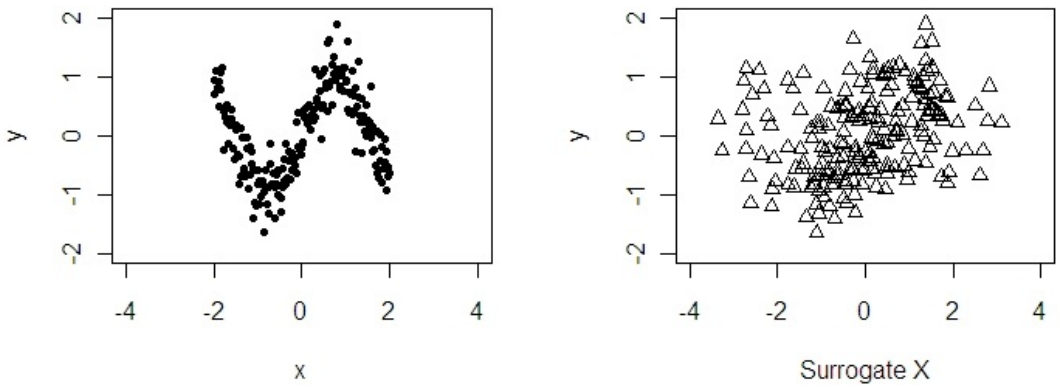

Figure 1.2: Measurement Error: Hidden Features.

1.2.2 Regression: The Details

Consider the scenario when a researcher is attempting to determine the effect of nitrogen level on the yield of a crop. To do this we choose randomly from several farms in the research region of interest. Since the level of nitrogen across several farms will vary, due to heterogeneity of soil, we can consider the levels of nitrogen as random. It is necessary to obtain a measurement of the nitrogen level in the soil which can be done by sampling the soil and performing a laboratory analysis on the sample. Clearly this is going to yield not only experimental error but also suffer from sampling error. We will assume a linear relationship between soil nitrogen and the crop yield.

We investigate the details here under the assumption that the covariate X is a random variable. Assume,

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 X_i + \epsilon_i \\ Z_i =& X_i + \delta_i \\ X_i \sim& N(\mu_x, \sigma_x^2) \\ \epsilon_i \sim& N(0, \sigma_\epsilon^2) \\ \delta_i \sim& N(0, \sigma_\delta^2), \end{aligned}$$

for all i. We can think of the error terms δ as resulting from sampling and laboratory analysis. Clearly the variables Y and Z are normal. Theory dictates that the joint distribution of (Y, Z) will be a bivariate normal with mean vector $\boldsymbol{\mu} = (\beta_0 + \beta_1\mu_x, \mu_x)$ and covariance matrix,

$$\Sigma_{yz} = \begin{bmatrix} \beta_1^2\sigma_x^2 + \sigma_\epsilon^2 & \beta_1\sigma_x^2 \\ \beta_1\sigma_x^2 & \sigma_x^2 + \sigma_\delta^2 \end{bmatrix}.$$

1.2.3 Classical Measurement Error

We assume that the surrogate variable Z is used in lieu of X and is related through the additive relationship, $Z = X + \delta$, where $E(\delta) = 0$ and $Var(\delta) = \sigma^2$. This results in a additive variability giving, $Var(Z) = Var(X) + Var(\delta) > Var(X)$. We make note that we are assuming that the variable X is a random variable. This classical measurement error model is what we used in the regression example of section 1.2.2.

1.2.4 Berkson Measurement Error

Unlike the classical measurement error model here we assume that the surrogate Z is related to the true X by the relationship, $X = Z + \delta$, where $E(\delta) = 0$ and $Var(\delta) = \sigma^2$. Here notice that the true values X have more variability than the observed surrogate. This is a result of the key difference from the classical model, the assumed additive relationship is reversed.

When to choose this model over the classical measurement error model is up for debate. Something worthy of mention is that although with the classical model, in the simple linear regression case with continuous X, we found that the estimator $\widehat{\beta}_1$, using the surrogate Z in lieu of X, was biased towards zero. Using the Berkson model does not result in this bias; however, this is not the case for using the surrogate variables as predictors in logistic regression. See Carroll (2006) and Reeves et al. (1998) for details.

1.2.5 Bayesian Methods

Measurement error in Poisson count regression can be encountered in several situations. The first being measurement error in the exposure. For example, suppose we are attempting to estimate the rate of occurrence of cancer for individuals exposed to a source of radiation, as was done in the LSS study. In this case recall bias is suspected for individuals report the amount of time they were exposed to the source

of radiation. Here the amount of time exposed to the treatment, radiation source, is then measured with error.

The second form of measurement error that can occur is simply an error related to the recording of the response variable. For example, it is often the case that a rape victim will not report the crime and so the true account of rapes within a region (time period) is greater than the reported count by law enforcement agencies. This is often the criticism criminologists has with the Universal Crime Reports (UCRs) that the Federal Bureau of Investigation publishes.

The third form of measurement error that can bias statistical inference is a measurement error in the regressors. We discussed the biases of this and measurement error in the response variable in section 1.2.2 for continuous variables. However, the effect is just as devastating for discrete data.

Many methods correcting for measurement error have been proposed from the two predominate statistical schools of thought. One such approach is from the Bayesian paradigm in which prior distributions are used on the variables subject to measurement error. Often times a training sample is performed in which a gold standard method is used to obtain the exact value of the variable while also measuring the variable with the error prone method. In this fashion an informative prior can be constructed and then incorporated into the model. This prior is can be chosen as the posterior distribution resulting from the training sample or a distribution approximating it, same mean and overall shape, with more variability. A posterior distribution is then calculated for the model using the data of real concern if it happens to be tractable which is a rather unrealistic expectation for complicated models. In the event that the posterior distribution cannot be derived explicitly, Monte Carlo methods (Monte Carlo Markov Chains - MCMC) are then implemented. See Ntzoufras (2009) for a detailed discussion on using WinBUGS and Gilks, Richardson, and Spiegelhalter (1996), Gamerman and Lopes (2006), and Robert and Casella

(2004) for details concerning MCMC methods. See Gustafson (2004) and Carroll (2006) for a discussion of measurement error models.

In many applications the researcher is looking to model a binary response variable as a function of several predictor variables. For example, we may wish to model recovery from cancer as a function of age, gender, and dose level of a particular drug. Commonly used regression techniques include probit or logistic regression. McGlothlin, Stamey, and Seaman (2008) propose a Bayesian method expanding on the work of Roy, Banerjee, and Maiti (2005) to handle situations of this nature when Measurement Error is present in the predictor variables.

1.2.6 Classical Methods

We provide here some Classical methods specifically designed for use when measurement error is suspected.

1.2.6.1 *Measurement error in exposure.* Consider the conditional mean function, $E(Y_i) = \lambda_i t_i$, where λ_i is the rate of occurrence, or intensity, per unit time and t_i is the length of the time period of exposure. Then the Poisson density is given as,

$$f(y_i|\lambda_i t_i) = \frac{e^{\lambda_i t_i}(\lambda_i t_i)^{y_i}}{y_i!}.$$

We call this the exposure model. Assume that the conditional intensity function is correctly specified. Now suppose that we are studying a system where we must rely on a respondents own recollection of how long they were subjected to a particular environmental hazard. Our outcome of interest is the onset of some disease. In situations such as these it is possible that the respondence will suffer from recall bias. For simplicity let us assume that the measurement error is uncorrelated with the covariates, $\mathbf{x}_i$, chosen for the conditional intensity.

One common approach to account for this fallible measurement in the exposure time is to consider t_i as a random variable from a gamma distribution. The resulting marginal distribution for the response variable, Y_i, is Negative-Binomial. Hence, a possible explanation for overdispersion is the presence of measurement error in the exposure time. Of course, we could choose any distribution for t_i and Cameron and Trivedi (1998) provide a derivation for the case of a general density $g(\cdot)$. After choosing the density, $g(\cdot)$, computational details are all that remain for the analyst.

1.2.6.2 *Additive measurement error in regressors.* At times it is reasonable to assume an additive gaussian error structure in the regressors. That is, $W = X + U$, where $U \sim N(0, \sigma^2)$. Now if one has access to replicated measurements on each regressor then we are able to obtain information about the moments of the error structure. In such cases, Carroll (2006) propose two methods, both of which are functional methods for generalized linear models. The first is called *conditional score* method and the second is called *corrected score* method. Both of these methods are computationally intensive and the reader is referred to the source for further information. Jordan et al. (1997) provide a similar method that is Bayesian and does not require replicated data.

1.2.6.3 *Multiplicative measurement error in regressors.* Multiplicative and additive error structures in the exponential conditional mean function can be shown to be algebraically similar. To see this assume, $\mathbf{w} = \mathbf{x} + \mathbf{u}$, where we assume some density for $\mathbf{u}$.. The conditional mean for the additive error structure is given as,

$$\begin{aligned} exp(\mathbf{w}'\boldsymbol{\beta}) =& exp((\mathbf{x} + \mathbf{u})'\boldsymbol{\beta}) \\ =& exp(\mathbf{x}'\boldsymbol{\beta}) exp(\mathbf{u}'\boldsymbol{\beta}) \\ =& exp(\mathbf{x}'\boldsymbol{\beta})\eta, \end{aligned}$$

where, $\eta = exp(\mathbf{u}'\boldsymbol{\beta})$.

Now consider the multiplicative measurement error structure,

$$exp(\mathbf{X}'\boldsymbol{\beta})\omega = exp(\mathbf{X}'\boldsymbol{\beta} + \epsilon),$$

where $\epsilon = ln(\omega)$. Clearly, adding another coefficient, $\beta_\epsilon = 1$, to the vector $\boldsymbol{\beta}$ will provide an additive measurement error structure.

The effect of both additive and multiplicative errors are also algebraically similar to a situation in which omitted regressors are important to the model. To see this let the conditional mean be dependent upon the vectors $\mathbf{x}$ and $\mathbf{z}$. Now suppose we include the vector $\mathbf{x}$ in the model but not the vector $\mathbf{z}$. Then the true conditional mean is given as,

$$\begin{aligned} exp(\mathbf{x}'\boldsymbol{\beta} + \mathbf{z}'\boldsymbol{\gamma}) &= exp(\mathbf{x}'\boldsymbol{\beta})exp(\mathbf{z}'\boldsymbol{\gamma}) \\ &= exp(\mathbf{x}'\boldsymbol{\beta})\omega_{\mathbf{z}}, \end{aligned}$$

where $\omega_{\mathbf{z}} = exp(\mathbf{z}'\boldsymbol{\gamma})$. Since we failed to include the vector $\mathbf{z}$ we may consider the effect of $\mathbf{z}$ as an unobserved heterogeneity effect which could be interpreted as measurement error in the regressors $\mathbf{x}$.

Hence, the effect of additive and multiplicative measurement error are algebraically the same. Furthermore, qualitatively the effects are similar to the omission of regressors important to the conditional mean. The difference lies in the interpretation of the error structure and how the error structure is to be modeled. For example, additive error structure with mean zero would correspond to a multiplicative error structure with mean unity.

1.2.6.4 *Measurement error in counts.* As briefly mentioned before the FBI UCRs are often criticized for being fallible. One reason is that reporting to the UCRs is voluntary and as such the data from a law enforcement agency may go unreported for several months. This is often the case for agencies operating in small rural communities. Furthermore, it is well documented that certain crimes often go

unreported and so the reported counts are often inaccurate and the true count is higher than the observed count.

The need for methods to account for underreporting is not simply limited to the applications of crime data. Examples of research using methods to account for underreporting can be found in applications from several different disciplines and research areas. The following applications are just a few notable examples: absenteeism in the workplace (Bramby, Orme, and Treble 1991; Johansson and Palme 1996), reporting of industrial accidents (Russer, 1991), safety violations in nuclear power facilities (Feinstein 1989; Feinstein 1990), criminal victimization (Feinberg 1981; Schneider 1981; Yannaros 1993), hospital medicine (Watermann, Jankowski, and Madan, 1994), and earthquakes and cyclones (Solow, 1993).

One simple method for correcting for fallible, inflated, counts is to assume that the True count, Y^t, is Poisson distributed with rate parameter λ_t (Cameron and Trivedi, 1998). Now let Y^o be the observed fallible count and assume that ϵ is Poisson distributed with rate λ_ϵ. This gives the observed count, $y^o = y^t + \epsilon$, a Poisson distribution with rate $\lambda_t + \lambda_\epsilon$. However, considering the fact that, $\epsilon \in \mathbb{Z}^+$, the observed count will always be larger than the truth and as such this model is of no use to us in the underreported scenario.

One method is to consider situations where ϵ can take on negative values, however, since the observed count is restricted to nonnegative integers we are forced to place restraints on the support of ϵ's distribution conditional on, $Y^t = y^t$. It would be unrealistic in this event to believe that Y^t and ϵ are independent and therefore a correlation structure would need to be considered. In this case a joint distribution for y^t and ϵ would be constructed and then the resulting distribution of y^o could be derived.

Binomial thinning is another method for modeling errors in counts (Cameron and Trivedi, 1998). As its name suggests it is only useful for deflating counts and

useful only for underreported counts. This mechanism operates on an event, conditional the event occurs, with constant probability that the event will go unreported. Thus, only true events are affected. That is, a Bernoulli process is responsible for determining if the count is reported or not.

Let $Y^t \sim Poisson(\lambda_t)$ and let $\pi = P(\text{event is observed})$. Then the distribution of the observed counts, Y^o, is Poisson with rate parameter $\pi\lambda_t$. Clearly this is a downward bias in the rate parameter λ_t, the parameter often of interest, since $0 \leq \pi \leq 1$.

We propose a method using our model, discussed later, by combining both a Poisson over count and a Binomial thinning to accommodate for over/under counts. Cameron and Trivedi (1998) makes the claim that models of this nature, methods incorporating a nonnegative only and a nonpositive only corrections, are underdeveloped and so this provides a large area of potential research.

1.3 Misclassification

Many obstacles pose severe problems when attempting to analyze count data and attempting to find relationships with the outcome to some set of covariates. Visibility bias (under counting), misclassification, and error measurements in the covariates are all common problems with poisson regression in a non-ideal environment: the typical social scientists realm of research. Misclassification also poses severe problems in medical, epidemiological, and the environmental sciences. Failing to account for misclassification can lead to biased estimates and underestimation of standard errors which can result in over stating the risks of certain covariates, or under stating them, leading to erroneous conclusions at best wasting valuable public resources.

In the following sections we consider the problem of misclassification for poisson response variables and provide a literature review for this topic.

1.3.1 Kutran 1975

Kutran (1975) proposed a Bayesian method to correct for visibility bias. Let A be the known area of the sample region, X be the number of individuals observed in this region, Y be the number of unobserved individuals in this region, and let $T = X + Y$ be the total individuals in the study region.

Now assume that x, y, and t represents the visible count, the non-visible count, and the total count from a training sample obtained from a region with area a, where a complete census was taken. Let λ be the population density, p be the visibility bias parameter, and assume that,

$$\begin{aligned} T|(\lambda, p) = T|(\lambda) &\sim pois(\lambda A), \\ X|(T, P, \lambda) = X|(T, p) &\sim Bin(T, p), \\ \lambda|p = \lambda &\sim \Gamma(\nu, a), \\ p|\lambda = p &\sim Beta(\alpha, \beta), \end{aligned}$$

where $\alpha = x + 1$, $\beta = y + 1$, and $\nu = t + 1$. Here we see that the observed count X is dependent on the total count and the visibility bias. Kutan then applies this procedure to a bicycle counting problem and to counting Gallinule Nests.

1.3.2 Stamey 2000

Stamey (2000) considers an extension of Kutran's (1975) work. Stamey extends the model to allow for the addition of misclassified terms considered to be false positives whereas Kutran stopped after accounting for false negatives. We define a *false positive* as a reported count that is not, in truth, an event of interest. Furthermore, a *false negative* is an event of interest, in truth, that is *not* counted as an event of interest.

Let A be the known area of the study region. Let T be the true number of the events of interest, Y be the number of false positives, and X be the number of false negatives. Then $Z \equiv T + Y - X$ is the number of reported events of interest. The

unobservable variables T, Y, and X are given the following distributions assuming that X and Y are independent:

$$T|\lambda \sim Poisson(\lambda A)$$
$$Y|\phi \sim Poisson(\phi A)$$
$$X|(t,\theta) \sim Binomial(t,\theta),$$

where θ is the probability of a false negative and ϕ is the rate at which false positives are reported. Stamey then shows that the marginal distribution of the reported events of interest is given as,

$$Z|(\lambda,\phi,\theta) \sim Poisson(A\lambda(1-\theta)+A\phi).$$

The Bayesian approach proposed by (Stamey, 2000) uses conjugate priors for λ, ϕ, and θ. Let us assume that t_0 represents the true number of events of interest in an area of A_0. Let y_0 be the number of false positives and x_0 be the number of false negatives. As described above we assume that $T|\lambda$ and $Y|\phi$ are Poisson random variables whereas $X|(t,\theta)$ is Binomial. Applying conjugacy for the priors we get,

$$\lambda \sim Gamma(t_0+1, A_0)$$
$$\phi \sim Gamma(y_0+1, S_0)$$
$$\theta \sim Beta(x_0+1, v_0-x_0+1),$$

where S_0 is used in the prior for ϕ in lieu of A_0 and v_0 in the prior for θ in lue of t_0 to reflect the fact it may come from a different source; e.g., expert opinion or a double sample. These priors are denoted as $g(\lambda)$, $g(\phi)$, and $g(\theta)$ respectively. As in the traditional interpretation we may think of x_0 to be the prior number of observed false negatives from a prior experiment or training sample of size v_0 making the use of prior data extremely intuitive.

1.3.3 Stamey, Young, Seaman (2007)

Applying techniques from McInturff et al. (2004) and Paulino et al. (2003) to both Whittemore and Gong (1991) and Sposto et al. (1992) Stamey, et al. (2007) provide a new Bayesian approach to Poisson regression where misclassification poses a problem. They do no rely on asymptotic distributions or assume that the misclassification parameters are known. The model is shown to perform well under simulation and is then applied to Atomic Bomb Survivor Data from Hiroshima and Nagasaki. The model is also applied to a respiratory tract infection example.

1.3.4 Classical Corrections

Whittemore and Gong (1991) propose a classical solution where maximum likelihood methods are used to obtain estimates. Their method is similar to that of Kutran's (1975) in that a training sample is conducted using a gold standard and the misclassification parameter p (visibility bias) is estimated using a binomial distribution.

The procedure is then applied to cervical cancer rates in major European countries to ascertain if the observed rate differences are due to misclassification or to actual cancer rates.

There have been of course many classical solutions for misclassification. One of which is proposed by Sposto, Preston, Shimizu, and Mabuchi (1992) which extends the results of Whittemore and Gong (1991) to allow for inferences across two groups applying their method to mortality rate estimates from the Hiroshima and Nagasaki survival data (LSS). A simple search in the literature will turn up many more.

1.4 Disease Mapping Basics

Consider a geographical area subdivided into multiple regions not necessarily of equal size or density. Now assume that for region i we have counts y_i of disease/crime occurrence from a total population n_i which is at risk of contracting

the disease/crime. The standard mortality ratio (SMR) is simply the ratio of the observed counts y_i to the expected number of counts E_i which as we will discuss can be unstable for small expected counts. Various ways of calculating the expected counts are available. Two popular methods are *external standardization* and *internal standardization*.

External standarization makes use of an existing standard table of stratified reference rates, r_j, for the disease in question. Then the expected counts are calculated by stratifying the population and summing the product of the rate and number of people at risk in stratum j. That is, $E_i = \sum_j n_{ij} r_j$ where n_{ij} is the total number of people at risk in stratum j of region i. However, the use of this procedure requires the availability of reference rates which may or may not be feasible.

Internal standardization uses the observed data to estimate the overall disease rate for the geographical area in question under a null hypothesis of constant disease risk. The expected count for each region i is simply the product of the number of people at risk in region i and the overall disease risk. That is, $E_i = n_i \overline{r} = n_i \frac{\sum_i y_i}{\sum_i n_i}$. Since reference rates may not be available in some cases, internal standardization is not only the simplest method to obtain expected counts but also may be the only way.

Once the expected counts have been calculated, we are faced with the task of estimating the disease risk θ_i for each region i. When attempting to estimate the disease risk within certain regions we might think to use the estimator $\frac{o_i}{e_i}$, commonly called the Standard Mortality Ratio (SMR), or the estimator $\frac{o_i}{t_i}$ for disease rate where o_i, e_i, and t_i represent the observed, expected, and total at risk for the i^{th} region. One major disadvantage to using these estimators is what is known as the *small number problem*; the estimators become severely unstable when either the expected number of incidents or total population at risk is small relative to the observed count. In the example of habitat burglary the same residence can be

burglarized multiple times within the time frame under consideration. In these cases the estimators could report an inflated estimate. Furthermore, when dealing with a very rare disease we might not observe even a single incident in a certain area thereby giving an estimate of zero, which is very unrealistic for a known disease. We refer the reader to Clayton and Kaldor (1987) for more information on why the standard maximum likelihood estimator of disease risk, the SMR, is not necessarily the best estimator.

Several methods have been proposed to circumvent the small number problem. One such possible solution is to pool counts from neighboring regions in an effort to increase the expected count. However, a tactic like this removes much of the spatial resolution in the data and we are left with less geographical information. Another possible solution is to employ a spatial smoother, a technique similar to that of weighted moving averages used in time series. We refer the interested reader to the plethora of literature concerning spatial smoothers.

1.5 Subset Selection

For completeness we provide here a brief review of Bratcher and Bhalla's (1974) subset selection procedure.

It is often the interest of a researcher to choose the best mean or rate parameter from several populations. A researcher might also be entrusted to make a decision between several possible choices and need a formal approach upon how to best accomplish this task. Bratcher and Bhalla (1974) proposed a subset selection procedure using a constant loss function and probabilities to make a decision between several possible options. We review this procedure now and will use this approach later for a few examples.

Let $\theta = \{\theta_i\}_{i \in \Omega}$ be the set of possible parameters of interest where Ω is an finite indexing set. An example for motivation might be a set of possible Poisson

rates, normal means, or binomial probabilities. Let θ_{max} be the parameter that we desire. Our objective is to provide a formal approach for constructing a subset of the θ_i's which we believe to be the most likely values of θ_{max}. Since we are required to make a decision, we will exclude the null set from being a possible choice; we have a total of $2^n - 1$ possible subsets to choose from where $n =\mid \Omega \mid$.

Each of these subsets is a possible choice for what we will call the *superior set* **S**. That is, we will choose a subset of the θ_i's to be our superior set **S**. When considered as a collection of decisions we have a total of n two decisions problems: include θ_i or exclude θ_i. From a notational standpoint we have the following decisions pertaining to each θ_i:

$$d^i_+ : \theta_i \in S \text{ and } d^i_- : \theta_i \in S^c.$$

We now adopt the constant loss function L defined as:
for the decision to include θ_i

$$L^i_+(\theta_i) = c_1 I[\theta_i \neq \theta_{max}]$$

and for the decision to exclude θ_i

$$L^i_-(\theta_i) = c_2 I[\theta_i = \theta_{max}],$$

where $I[\cdot]$ is the indicator function which has value 1 or 0 depending on the truth of the argument.

Clearly we have no way of knowing which θ_i is θ_{max} in practice and as such our goal is to come up with some method upon which to choose the superior set **S** in such a fashion as to minimize risk. To that end we note that once we have chosen a superior set **S** we have made a total of m inclusions where $1 \leq m \leq n$ and as such we have incurred a total loss of $m * c_1 + c_2$ if we failed to include θ_{max} and $(m-1)c_1$ if we included θ_{max}.

With the consideration that our objective is to choose a subset which contains θ_{max}, albeit we want the smallest possible such subset, we will commit a more severe

error if we fail to include θ_{max} in the superior set $\mathbf{S}$ than if we include θ_{max} with other θ_i's. As such we shall choose c_2 to be greater than c_1.

It should be clear that the decision to include or not to include the value θ_i in the superior set should depend in some way the losses that will result from the decision we make. It is desirable to make the choice that would result in the smallest loss. However, since we do not know which θ_i is the value θ_{max} we do not know which decision will result in the smallest loss. Obviously, if we knew the value of θ_{max} we wouldn't even be considering other values θ_i in practice. So, we make the decision corresponding to the smallest expected loss, otherwise known as Risk. We let x be a vector or sufficient statistics calculated from collected data. Since we want to minimize the risk we will choose to include θ_i provided,

$$E(L_+^i(\theta_i|\mathbf{x})) \leq E(L_-^i(\theta_i|\mathbf{x})).$$

Writting the above expectations as functions of $P(\theta_i = \theta_{max})$ gives us,

$$E(L_+^i(\theta_i|\mathbf{x})) = 0 * P(\theta_i = \theta_{max}|\mathbf{x}) + c_1 * P(\theta_i \neq \theta_{max}|\mathbf{x}),$$

and,

$$E(L_-^i(\theta_i|\mathbf{x})) = 0 * P(\theta_i \neq \theta_{max}|\mathbf{x}) + c_2 * P(\theta_i = \theta_{max}|\mathbf{x}).$$

With some algebra we then obtain,

$$E(L_+^i(\theta_i|\mathbf{x})) = c_1 * (1 - P(\theta_i = \theta_{max}|\mathbf{x})),$$

and,

$$E(L_-^i(\theta_i|\mathbf{x})) = c_2 * P(\theta_i = \theta_{max}|\mathbf{x}).$$

Since we wish to have $E(L_+^i(\theta_i|\mathbf{x})) \leq E(L_-^i(\theta_i|\mathbf{x}))$ as our rule to include θ_i, our decision is based on the relation,

$$c_1 * (1 - P(\theta_i = \theta_{max}|\mathbf{x})) \leq E(L_-^i(\theta_i|\mathbf{x})) = c_2 * P(\theta_i = \theta_{max}|\mathbf{x}),$$

which reduces to,

$$P(\theta_i = \theta_{max}|\mathbf{x}) \leq \frac{1}{1+\frac{c_2}{c_1}} = \frac{1}{c+1},$$

where we have chosen $c = \frac{c_2}{c_1}$. With this consideration we do not need to specify each of the individual losses c_1 and c_2 but rather simply the ratio c.

We have now produced a formal approach to selecting various values for a parameter of interest while minimizing the risk associated with such a decision. Our rule for including a possible θ_i is $P(\theta_i = \theta_{max}|\mathbf{x}) \leq \frac{1}{c+1}$.

1.6 The Model

1.6.1 A Conjugate Hierarchical Model

For the purpose of this dissertation we will use the following parametrization of the gamma pdf,

$$f(t|(\alpha,\beta)) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} exp(-t\beta). \tag{1.14}$$

In the course of this work we shall be using variations of the following hierarchical model. First let y_i be the observed count for the i^{th} unit under study and assume the structure,

$$\begin{aligned}
Y_i|\lambda_i &\sim Poisson(E_i\lambda_i\eta) \qquad (1.15)\\
\eta|(a,b) &\sim Beta(a,b)\\
\lambda_i|(\mu_i,\alpha_i) &\sim Gamma(\mu_i\alpha_i,\alpha_i)\\
\alpha_i|\nu &\sim \chi^2(\nu)\\
\mu_i &= \gamma_i\kappa\\
log(\gamma_i) &= \mathbf{x}\boldsymbol{\beta}\\
\boldsymbol{\beta} &\sim h(\cdot),
\end{aligned}$$

where E_i is the expected number of events for subject i, λ_i is the inflation factor for subject i, $0 < \eta \leq 1$ is a parameter used to correct for under-counts, κ is used to correct for over-counts and can possibly be dependent upon covariates, $h(\cdot)$ is a probability distribution, and ν, a, and b are fixed. Notice that this means λ_i is a multiplicative factor able to adjust the subject rate away from the expected E_i, which we call the *inflation factor*. Secondly, note that the mean of the Gamma distribution is μ_i which results in a log-linear model on the mean of the inflation factors λ_i rather than the Poisson rates themselves. Furthermore, through the use of the η_i and κ parameters we can adjust the model to incorporate misclassification/measurement error (over and under counts) in the response variable. We make note that although we include a parameter to correct for over-counts, namely κ, we do not analyze a data set of such nature in this dissertation. We merely include this parameter for to demonstrate the flexibility of this model for researchers. This model is an extension of the one originally proposed by McBride (2006) where we have added the η_i and κ parameters to correct for over and under counts in the data. If both η_i and κ are equal to unity for all i then we would have the model proposed by McBride.

Several methods are available for corrections of measurement error in the covariates. One such possible method would be to place priors on the components of $\mathbf{x}$, considering the x_i as random variables, and use a training sample scheme to construct a posterior distribution for the x_i which would be used to correct the observed counts. The point is, the researcher is free to use reliable methods for such corrections when certain covariates are suspected to be measured with error.

The expected counts E_i can be obtained via internal or external standardization as discussed in detail in section 1.4. In the spatial example presented in this dissertation we will calculate the E_i via internal standardization under the hypothesis that the entire region has an overall constant disease rate. This is done as,

$$E_i = n_i \overline{r} = n_i \frac{\sum y_i}{\sum n_i},$$

where n_i is the number of subjects at risk for region i and $\overline{r}$ may be thought of as an overall disease risk. The example using baseball homerun data does not use an expected count and will be discussed in detail in section 2.2.

As mentioned, often times situations arise in count data where over and under dispersion is present in the data. It is well known that although consistency may hold for the estimator if the conditional mean is specified correctly the reported standard errors may be overly optimistic leading to erroneous conclusions at best (Cameron and Trivedi, 1998). Many efforts have been made to account for this including quasi-likelihood methods, Generalized Count Regression, using a Negative Binomial model instead of Poisson, and several Bayesian methods including Albert's (1992) model which mixes quasi likelihoods with Bayesian methods.

In our model we allow for the Poisson rates λ_i to vary about a mean $E(\lambda_i) = \mu_i$. This allows for the data to be over/under dispersed naturally. In addition to the simple problem of over/under dispersion the concern that we have correctly specified the conditional mean is relaxed. That is, we are able to incorporate covariates of interest into the model, provide inference, while still allowing for the heterogeneity that would be present in the absence of other significant and unknown or unmeasurable covariates. This is advantageous when we are interested in knowing if certain covariates are useful predictors but realize that there may be other sources of variation for which we have not yet been able to accommodate, a common situation in public health data and criminology research. This added error structure is then used to correct for the missing covariates.

Our ability to control for these unobserved sources of variation is found in the α, η, and κ parameters. The α parameter adjusts the variability of the gamma distribution and, hence, is a measure of our confidence in the log-linear model.

When we have reason to believe that we have not accounted for all the sources of variation we can simply change ν to adjust the variance of the Gamma distribution accommodating the overdispersion suspected in the data. The other two parameters allow the researchers to control for under and over counts, ν and κ respectively, when such problems are suspected to exist within the data.

Motivation for where we place the over and under count correction parameters stems from how we envision the statistician collecting expert opinion and prior information from the researchers. In order for an event to go unreported the event must have already occurred and, hence, we chose to use a parameter effecting the rate of the Poisson distribution directly. An inflation of counts resulting from over reporting or false reports do not have the same requirement that the event did, or did not, occur and so we chose to place the parameter as a multiplicative effect on the mean of the inflation factors. Although we have listed these parameters as not being dependent upon covariates such an adjustment would be relatively easily allowing for the researchers to include a covariate suspected to influence the reporting of these events.

1.6.2 Variance Relationship with ν

As we will discuss further in later chapters the choice of ν is extremely important and can have a significant impact on the conclusions made by the researchers concerning the observables relationship with the chosen regression parameters. What we provide here are mathematical relationships of the observables variance and the choice of ν. First, we note that for $X \sim \chi^2(\nu)$, $E\left(\frac{1}{X}\right) = \frac{1}{\nu-2}$.

Now notice that for model (1.15) we have $E\left(Y_i|(\lambda_i, E_i)\right) = E_i\lambda_i$. Using the double expectation formula, where we omit the conditioning on μ_i for notational convenience, we have,

$$
\begin{aligned}
Var(Y_i|E_i) =& E(Var(Y_i|\lambda_i, E_i)) + Var(E(Y_i|\lambda_i, E_i)) \\
=& E(E_i\lambda_i) + Var(E_i\lambda_i) \\
=& E_i\mu_i + E_i^2\left(E\left(Var\left(\lambda_i|\alpha_i\right)\right) + Var(E(\lambda_i|\alpha_i))\right) \\
=& E_i\mu_i + E_i^2\left(E\left(\frac{\mu_i}{\alpha_i}\right) + Var(\mu_i)\right) \\
=& E_i\mu_i\left(1 + \frac{E_i}{\nu - 2}\right).
\end{aligned}
$$

We can think of the quantity $\left(1 + \frac{E_i}{\nu-2}\right)$ as a scale parameter similar to (1.9). Notice that as ν gets asymptotically large the variance in the observables, Y_i, approaches their mean, $E_i\mu_i$, satisfying the assumption of the Poisson distribution. An interpretation then for ν is a measure of our confidence in the observables following a Poisson distribution; that is, large values place more confidence on this assumption. Furthermore, for small inflations of the variance above the mean, as expected if the data were approximately Poisson, we would want to choose $\nu - 2$ to be close to E_i.

CHAPTER TWO

Homerun Modeling

In the sport of Baseball very few events rival that of a homerun, with the exception of the very rare triple play. With a single hit the outcome of the game can change and bring fans to their feet. Baseball fans are often attracted to the powerhouse hitters such as Babe Ruth, Micky Mantel, and other well known players.

In this section we apply model (2.1) below to homerun rates for Marc McGuire, Sammy Sosa, and Berry Bonds, previously considered by McBride (2006). Here we expand on McBride's analysis providing new plots and a comparison of the developed Bayesian model with popular classical models designed for similar use.

2.1 Mathematical Model

Recognizing that there will not be under and over counts for this data set, for a fixed player we define Y_i be the observed count for year i and assume the following hierarchical structure,

$$Y_i|\lambda_i \sim Poisson(t_i\lambda_i) \tag{2.1}$$

$$\lambda_i|(\mu_i, \alpha_i) \sim Gamma(\mu_i\alpha_i, \alpha_i)$$

$$\alpha_i|\nu \sim \chi^2(\nu)$$

$$log(\mu_i) = \mathbf{x}\boldsymbol{\beta}$$

$$\beta_i \sim h_i(\cdot),$$

where t_i is the number of at-bats for year i, λ_i is the rate of homeruns for year i, $h_i(\cdot)$ is a uniform density, and ν is fixed.

Albert's (1992) model and model (2.1) attempt to model the same type of data just for a different set of players. Albert (1992) models the Poisson rates directly via

the log link and uses quasi-likelihood methods to correct for overdispersion. Since we are considering the homerun rate as a random variable from a gamma distribution, we are modeling the mean of the homerun rates rather than the rates themselves.

A key difference in the interpretation between model (2.1) and (1.15) is that we are not using t_i as the expected number of homeruns. Rather t_i is the total number of at-bats for the given player in year i. However, we are modeling the mean of the λ_i's as done in model (1.15). Under the consideration that homeruns are actually not at all common events in baseball, it is not unreasonable think of λ_i as an approximation to the percentage of the at-bats that will result in a homerun although this is not exactly true. As a result of this consideration we may consider our log-linear model on the mean of the homerun rates as opposed to the mean of *inflation factors*. We make note here that we could have chosen to take the players average homerun rate, internal standardization with the assumption of a constant rate, and used that as an *expected* count resulting in the interpretation found in model (1.15). This shows, depending on how we use the parameter E_i in model (1.15), the interpretations of the resulting model can be flexible.

2.2 The Analysis

In the context of this problem our parameter vector of interest is $\boldsymbol{\lambda}_j$, the individual homerun rates for each of the j players in question. To make comparisons between the players we could use a classical approach and compare confidence intervals. Albert (1992) takes the Bayesian approach and compares posterior probability intervals. In this analysis we use the subset selection procedure developed by Bratcher and Bhalla (1974), used by Hamilton et al. (2008) and Stamey et al. (2004), to determine which of the three players were most likely, based on probabilities, the better homerun hitter.

2.2.1 Problem Set Up

It is generally believed that a player matures to a peak near the middle of his career and then declines in performance until retirement. Such performance could be modeled using a quadratic form. This quadratic like behavior is evident in Sammy Sosa's performance over his career of playing Baseball shown in Figure 2.1. The last five years show a decreasing trend in his homerun hitting rate. As for both Berry Bonds and Mark McGuire it appears that the quadratic behavior is also present but not fully realized due to lack of end career data. Mark McGuire retired early and his final year was near the peak of his career while Berry Bonds shows some decline in his final few years.

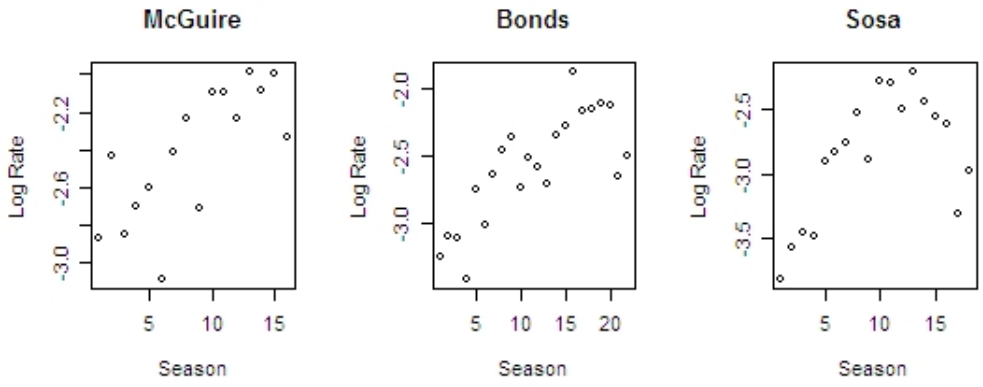

Figure 2.1: Observed Log HR Rates

The model we use is structured as model (2.1) where we place the log linear relationship on the rate means with the season of play and place uniform priors on the coefficient terms. That is,

$$log(\mu_i) = \beta_{0j} + \beta_{1j}(i - \bar{i}) + \beta_{2j}(i - \bar{i})^2, \tag{2.2}$$

$$\beta_{ij} \sim Uniform(a, b),$$

where a and b are specified and λ_{ij} is the homerun rate for the i^{th} season for player j. Two other hierarchical models proposed are Albert (1992) and (2000) which

proposes a hierarchical GLM. We refer the interested reader for further information to these sources.

In order to develop an understanding for a prior structure for each of the β terms consider the following. We expect a player to have a homerun rate near zero (say .01) near the beginning of the career and for some large value of i. During the *productive* years of a player's career we expect to see a homerun rate of no more than 0.17. With the assumption of a 20 year career, setting up a system of equations and solving for the beta terms gives us $\beta_0 = -6.0$, $\beta_1 = 0.82$, and $\beta_2 = -0.041$. Figure 2.2.1 displays this expected behavior. Hence, we choose uniform priors about these values for our analysis.

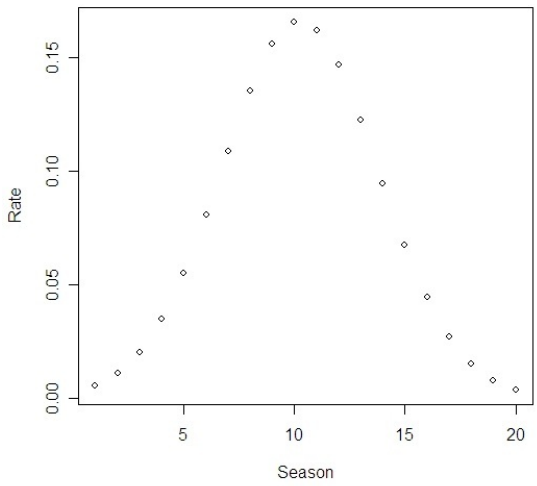

Figure 2.2: Prior Belief of Expected Rates

2.2.2 Analysis with $\nu = 10$

For the purposes of clarification we make note that the analysis using $\nu = 10$ does not provide a resulting model where comfortable inference can be made. That is to say, $\nu = 10$ does not seem to provide good results for the observed data. This analysis is placed here to demonstrate that improper choices for ν can lead to erroneous conclusions due to inflated posterior variability.

Using model (2.1) we use the program WinBUGS to estimate the posterior distribution. This program is widely used and available free, as of the writing of this dissertation, to any who wishes to acquire the program. WinBUGS uses Monte-Carlo Markov chain (MCMC) methods to estimate the posterior distribution. With the consideration of the expected behavior as discussed in section 2.2.1 we set uniform priors of Uniform(-10,0), Uniform(-3,5), Uniform(-4,1) for the terms β_0, β_1, and β_2, respectively. We allow for some probability weight on positive values for the quadratic term in the event that a player does not follow a decreasing trend in his later seasons. Lastly a chi-square distribution with $\nu = 10$ for the shape parameter α of the gamma distributed λ_i's is used.

We note that we have place relatively non-informative priors on the beta terms. Unfortunately, this induces an informative prior on the μ_i's for early seasons; however, as the seasons increase, the induced priors become relatively non-informative. Since our objective is to choose which season was the best season for the players in question, and then to choose which player had the best homerun hitting rate, we accept the consequence of this induced prior as these seasons will most likely occur in mid to late careers.

Three chains were run with a 30,000 burnin period using a thinner of 5 to reduce autocorrelation; total sample size 90,000. Convergence diagnostics were used to determine chain convergence and no issues were found: Gelman-Rubin statistic, autocorrelation plots, trace plots and density plots. We then generated 30,000 further iterations with a thinner of 5 to reduce autocorrelation. This resulted in an effective sample of 90,000 from the posterior distribution. The density estimates were smooth and bell shaped while the trace plots showed good mixing between the three chains. Table 2.1 provides posterior point estimates resulting from the generated chains.

Table 2.1: Baseball Parameter Point Estimates $\nu = 10$

Player	Parameter	Mean	SD	2.5%	5%	Median	95%	97.5%
Mcguire	Int.	-2.183	0.238	-2.679	-2.590	-2.172	-0.182	-1.750
	Linear	0.046	0.037	-0.025	-0.013	0.045	0.109	0.123
	Quadratic	-0.003	0.009	-0.021	-0.0178	-0.002	0.116	0.014
Bonds	Int.	-2.205	0.199	-2.618	-2.546	-2.197	-1.892	-1.837
	Linear	0.039	0.024	-0.007	1.27E-04	0.038	0.079	0.087
	Quadratic	-0.004	0.004	-0.012	-0.011	-0.004	0.003	0.004
Sosa	Int.	-2.246	0.233	-2.733	-2.645	-2.236	-1.886	-1.821
	Linear	0.040	0.036	-0.029	-0.018	0.040	0.100	0.113
	Quadratic	-0.012	0.007	-0.027	-0.025	-0.012	-6.21E-05	0.002

In order to evaluate the need for the quadratic model, credible sets are constructed for the covariate parameters and presented in Table 2.1. For completeness we have included credible sets for all terms in the model. Notice that for Mark McGuire the 95% credible set, as well as the 90% credible set, for the quadratic term includes zero. Although the quadratic behavior in the McGuire data is not evident the quadratic behavior of the Sosa data is quite evident but the 95% credible set for Sosa's quadratic term fails to exclude zero, although the upper bound is near zero, and the 90% credible set has an upper bound is practically zero. The Bonds data does show near quadratic behavior while the 95% credible set includes zero along with the 90% credible set, although the upper bounds are near zero.

The issue of obtaining a credible set for the quadratic term for Sammy Sosa that includes the value of zero when the plots of the observed data suggests that these terms should be non-zero leads us to believe that either the model does not fit the data well or that we have allowed for too much posterior variability, either through the support of our uniform priors or the choice of the ν parameter. In section 2.2.2.2 we rerun the analysis using other choices for the ν parameter and demonstrate that the choice of ν is crucial.

2.2.2.1 *Subset selection.* Until now we have focused on simply providing the reader with the standard point estimates and credible sets for the parameters of interest (namely the beta terms). However, our goal is to select which season was the peak season for the individual players and select from the players which had the highest homerun rate at the peak season. We use here the optimal subset selection procedure developed by Bratcher and Bhalla (1974) to answer the question at hand. A constant loss function is chosen with the ratio $c = 24$ giving rise to a probability cut off value of 0.04.

We present here the probabilities of each season being the best season for hitting homeruns by each player in Table 2.2. Using the subset selection procedure discussed previously, we would conclude that McGuires best seasons were either 8, 10, 11, 13, 14, or 15; Bond's best seasons were either 16, 19, or 20; Sosa's best seasons were either 10, 11, 13, or 14. However, as these values resulted from an improper choice of ν we do not recommend that these conclusions be made base on this model.

Table 2.2: PICs for each season $\nu = 10$

Season	McGuire	Bonds	Sosa	Season	McGuire	Bonds	Sosa
1	0.01311	0	5.56E-06	12	0.009156	0	0.01566
2	1.44E-04	0	0	13	0.2774	0	0.4994
3	0	0	0	14	0.08783	0.004022	0.04052
4	0	0	0	15	0.3009	0.004656	0.007522
5	0	0	0	16	0.01084	0.6444	0.00425
6	0	0	2.00E-04	17	na	0.03102	0
7	6.78E-04	1.11E-05	2.39E-04	18	na	0.03572	2.78E-05
8	0.1084	1.00E-04	0.01381	19	na	0.06259	na
9	0.003311	0.002211	5.56E-06	20	na	0.2147	na
10	0.1039	0	0.2283	21	na	5.56E-05	na
11	0.08436	4.44E-05	0.1901	22	na	4.56E-04	na

2.2.2.2 *Sensitivity analysis.* As illuded to in sections 1.6.2 and 2.2.2, mentioned by McBride 2006, and further discussed in section 3.2.2 the choice for ν

can greatly influence the posterior distributions variance and thereby influence the conclusions the researcher makes concerning the system under study. Therefore, considerable care must be taken into account when making the choice for ν. That is, too small values will greatly exaggerate over-dispersion increasing the posterior variance and lead to erroneous conclusions. Furthermore, for too large values of ν our credible sets may be overly optimistic and too tight leading researchers to falsely conclude statistical relationships where there is none.

Here we provide a sensitivity analysis, in Table 2.3, where we change the value of ν, leaving all other things equal, and report posterior mean, median, standard deviation, and the standard quantiles of interest for $\nu = 15, 20, 25$; for the value $\nu = 10$ the reader is refered to Table 2.1 above. To save space we abbreviate the payers' names by using the first letter of their last name; that is, M represents Mcguire, ect.

Figures 2.3, 2.4, and 2.5 display the same results but for $\nu = 5$ to 30. The noticeable trend is in the posterior standard deviations for the parameters; as ν gets large, the posterior standard deviation reduces. This trend is also noticed in section 3.2.2 where we comment on it further. Notice that the parameter point estimates seem relatively stable, the scales on the graphs are relatively tight, and the only real change is in the reduction of posterior variance resulting in tighter credible sets.

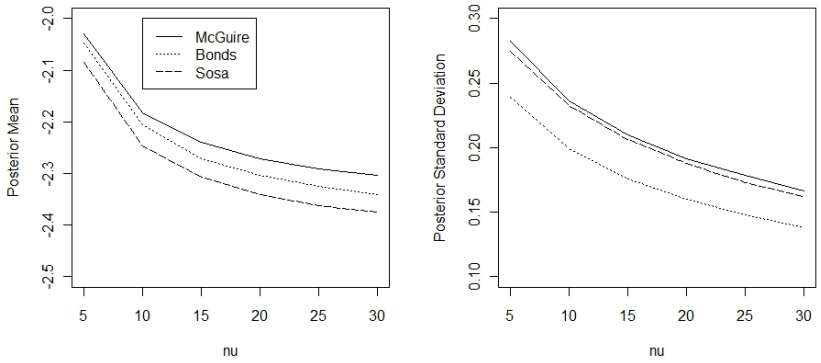

Figure 2.3: Baseball: Sensitivity Analysis for β_0.

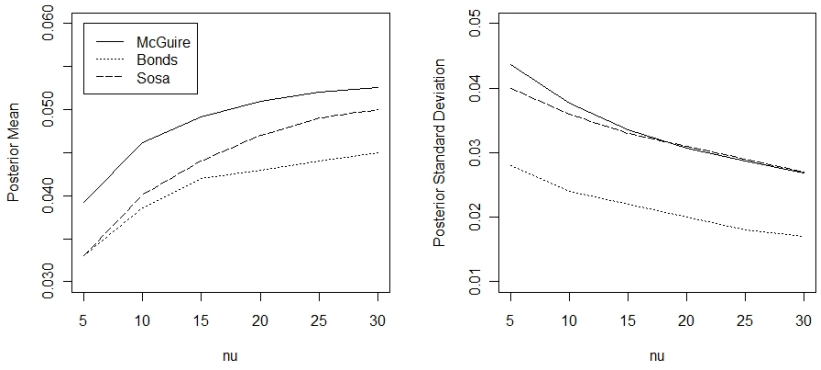

Figure 2.4: Baseball: Sensitivity Analysis for β_1.

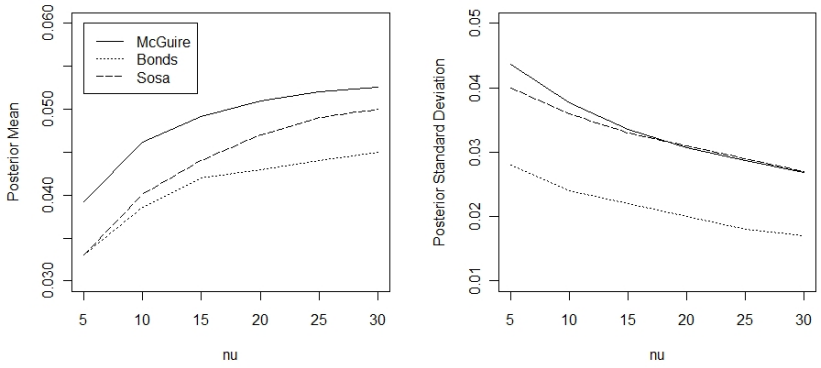

Figure 2.5: Baseball: Sensitivity Analysis for β_2.

2.2.3 Analysis with $\nu = 100$

For purposes of demonstration on choosing ν to be large we rerun the analysis here with $\nu = 100$. With the consideration of the expected behavior as discussed in section 2.2.1 we set uniform priors of Uniform(-10,0), Uniform(-3,5), Uniform(-4,1), same as for the previous analyses, for the terms β_0, β_1, and β_2 respectively. We allow for some probability weight on positive values for the quadratic in the event that a player does not follow this quadratic trend.

We note that we have place relatively non-informative priors on the beta terms. However, this induces an informative prior on the μ_i's for early seasons but as the seasons increase the induced priors become relatively non-informative. Since our objective is to choose which season was the best season for the players in question,

Table 2.3: Sensitivity Analysis of Posterior Point Estimates.

		$\nu = 15$		$\nu = 20$		$\nu = 25$	
	Covar.	Mean (SD)	(2.5% , 97.5%)	Mean (SD)	(2.5% , 97.5%)	Mean (SD)	(2.5% , 97.5%)
M.	β_0	-2.24 (0.21)	(-2.676, -1.855)	-2.271 (0.191)	(-2.667, -1.917)	-2.291 (0.178)	(-2.659, -1.96)
	β_1	0.049 (0.033)	(-0.014, 0.118)	0.051 (0.031)	(-0.007, 0.114)	0.052 (0.029)	(-0.002, 0.111)
	β_2	-0.003 (0.008)	(-0.019, 0.012)	-0.003 (0.007)	(-0.017, 0.011)	-0.003 (0.007)	(-0.017, 0.01)
B.	β_0	-2.308 (0.206)	(-2.74, -1.931)	-2.341 (0.188)	(-2.731, -1.995)	-2.362 (0.173)	(-2.719, -2.041)
	β_1	0.044 (0.033)	(-0.019, 0.111)	0.047 (0.031)	(-0.011, 0.11)	0.049 (0.029)	(-0.006, 0.109)
	β_2	-0.013 (0.007)	(-0.027, 0)	-0.014 (0.006)	(-0.027, -0.002)	-0.014 (0.006)	(-0.026, -0.003)
S.	β_0	-2.271 (0.176)	(-2.636, -1.945)	-2.303 (0.16)	(-2.632, -2.006)	-2.325 (0.148)	(-2.63, -2.047)
	β_1	0.042 (0.022)	(0, 0.085)	0.043 (0.02)	(0.005, 0.084)	0.044 (0.018)	(0.009, 0.082)
	β_2	-0.004 (0.004)	(-0.011, 0.003)	-0.004 (0.003)	(-0.011, 0.002)	-0.004 (0.003)	(-0.01, 0.002)

then further choosing which player had the best homerun hitting rate, we accept the consequence of this induced prior as these seasons will most likely occur in mid to late careers.

We used WinBugs to simulate data from the posterior values for the beta terms via Monte-Carlo Markov chain methods (MCMC). Two chains with a 10,000 burn in followed by 30,000 iterations from each chain after using a thinner of 10. Table 2.4 gives point estimates and quantiles for the beta terms. Figure 2.6 is plot of the estimated log rate curve by player while figure 2.7 gives a plot of the three estimated curves for comparison of the players.

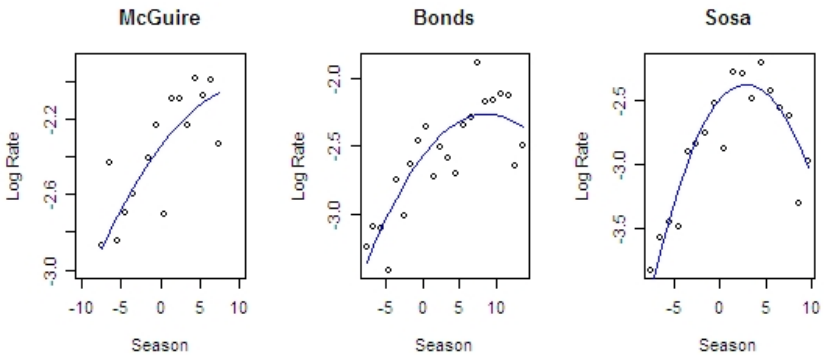

Figure 2.6: Baseball: Estimated Log Rate Curves

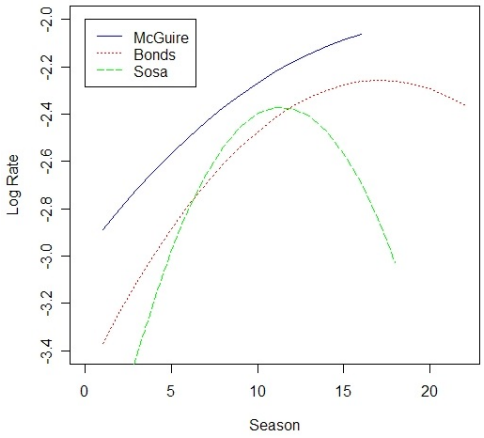

Figure 2.7: Baseball: Comparison of Players Estimated Curves

Table 2.4: Baseball Parameter Point Estimates $df = 100$

Player	Parameter	Mean	SD	2.5%	5%	Median	95%	97.5%
McGuire	β_0	-2.346	0.112	-2.573	-2.534	-2.344	-2.165	-2.131
	β_1	0.055	0.018	0.021	0.026	0.055	0.085	0.091
	β_2	-0.002	0.004	-0.011	-0.010	-0.002	0.005	0.006
Bonds	β_0	-2.392	0.090	-2.574	-2.543	-2.390	-2.246	-2.219
	β_1	0.048	0.012	0.026	0.029	0.048	0.067	0.071
	β_2	-0.004	0.010	-0.008	-0.007	-0.004	-0.001	-0.001
Sosa	β_0	-2.422	0.104	-2.634	-2.597	-2.42	-2.254	-2.224
	β_1	0.056	0.019	0.020	0.025	0.056	0.088	0.095
	β_2	-0.015	0.004	-0.023	-0.022	-0.015	-0.009	-0.008

2.2.4 Subset Selection

Our goal is to select which season was the peak season for the individual players and select from the players which had the highest homerun rate at the peak season. We use here the optimal subset selection procedure developed by Bratcher and Bhalla (1974) to answer the question at hand. A constant loss function is chosen with the ratio $c = 8$.

These three players brought much excitement to the game of Baseball during the final years of the twentieth century and the early few years of the twenty-first century. It is these years that the race for the title of homerun king became a topic of much conversation. So, for our considerations we will restrict our analysis to these years when determining which season was the best season for each player; contrasted from the analysis performed in section 2.2.2.1.

Table 2.5 table gives the probability of inclusion for each of these mentioned seasons. Using the subset selection procedure referenced, for our chosen constant $c = 8$ we would conclude that for Marc McGuire his peak season was his final season of play, 2001. Barry Bonds peak season was either 2001, 2002, 2003, or 2005. For Sammy Sosa we conclude his peak career was either 1999 or 2000.

Table 2.5: PIC For Peak Season

Year	McGuire	Bonds	Sosa
1995	0.013	NA	NA
1996	0.046	NA	0.000225
1997	0.075	NA	0.00102
1998	0.080	NA	0.0743
1999	0.070	0.063	0.489
2000	0.056	0.143	0.350
2001	0.659	0.187	0.069
2002	NA	0.161	0.0120
2003	NA	0.117	0.00470
2004	NA	0.080	NA
2005	NA	0.246	NA

To answer the question of, "Who had the highest homerun hitting rate at the peak of their career?" we apply the subset selection procedure. We obtain the probability of inclusion (PIC) for Mcguire to be 0.88, Bonds to be 0.10, and Sosa to be 0.012 (probabilities obtained with rounding). Therefore, we would conclude, using a constant loss of $c = 8$, that McGuire had the greatest peak in hitting homeruns.

2.3 Bayes-Classical Comparison

In what follows we make comparisons in the performance of model (2.1), with $\nu = 100$, to the frequentist approach to the problem. The choice of $\nu = 100$ is arbitrary and only for the purposes of comparison as no studies have yet been done on optimal choices for ν using model (1.15). We model the homerun rates directly using the log-link and we perform two analyses: the first we model as,

$$log(\lambda_i) = \beta_0 + \beta_1 s_i + \beta_2 s_i^2,$$

and the second we model as,

$$log(\lambda_i) = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 t_i,$$

where s_i and t_i represent i^{th} season and the number of at bats the player had in the i^{th} season. We note that for the more complicated model the *Akaike Information*

Criterion (AIC) values are considerably larger than for the simpler model. If we were to base our model choice on this value alone we would then choose to exclude the number of at bats a player had in a given season, which of course defies intuition considering the context of the problem. We first discus the results for the simpler model. Furthermore, the results presented consider all parameters included in the analysis, regardless of the achieved p-value, for purposes of comparison between the competing models.

Table 2.6: AIC's for Models Considered.

Model	Sosa	Bonds	Mcguire
At Bats	99.2256	71.9674	56.6030
Non At Bats	61.9327	61.4531	29.0311

2.3.1 Ignoring At Bats

Using PROC GENMOD in SAS 9.2 with the Poisson distribution and log-link we analyzed the number of homeruns directly (using the log link) as a function of both season and the square of season. Due to lack of fit we set the scale parameter to be estimated using the square root of the Pearson's Chi-Square statistic divided by it's degrees of freedom which is used to accommodate for overdispersion. As for the estimates of the coefficients we find similar results as found using model (2.1). Here we display the estimated classical curves, the observed data, the estimated number of homeruns for each player as obtained by model (2.1), and a comparison of the 95% confidence intervals and 95% credible set widths.

Notice that the classical approach models the trend in the mean of the homeruns as a function of season rather well. However, this model does not use the information available in the number of at bats a player has in a given season. The dashed line in in Figures 2.8, 2.9, and 2.10 shows the estimated number of homeruns for each player using model (2.1) and appears to be *over fitting* the data. This is merely an illusion.

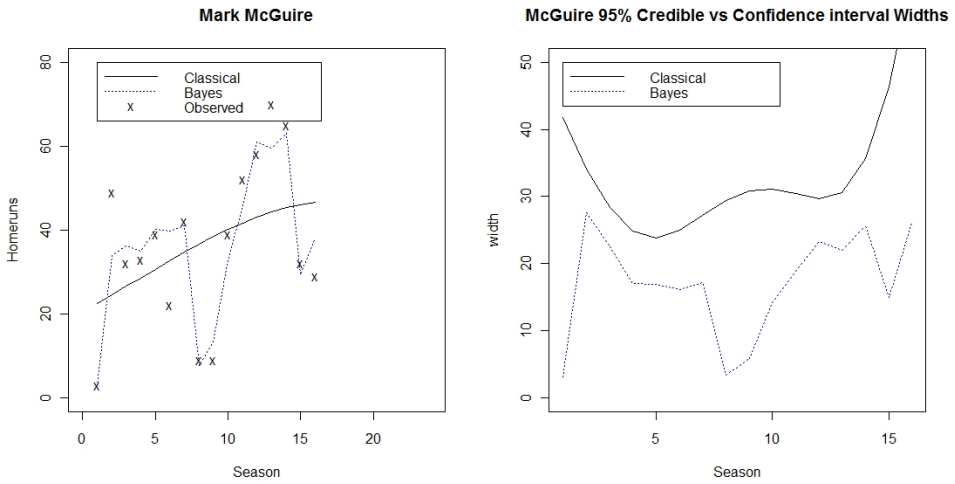

Figure 2.8: Baseball: Comparison of classical method and Bayesian method for McGuire.

We first remind the reader that the general linear model and model (2.1) are modeling differing responses. Model (2.1) models the mean rates of the players using a log-link while the GLM models the rate of the observed homeruns directly through the log-link. Furthermore, a difference in the Bayesian approach is that the rates are considered random variables versus the classical approach that they are fixed but unknown quantities. Second, using model (2.1), we construct the predicted number of homeruns a player had in a given season based on an estimate of the players rates multiplied by the number of at bats the player saw in a given season, something this GLM does not account for. This is why model (2.1) more closely follows the observed homeruns giving the appearance of *over fitting* the data.

If we were to make comparisons between players based on the GLM model it is worthy to note that we would conclude the peak of Sammy Sosa's career, his 10th through 14th seasons, appears to outperform Mark McGuire's peak, his 11th through 16th seasons. This is due to the fact that during these seasons Sammy Sosa did indeed hit more homeruns. However, the fact that the number of at bats for each player is not taken into account by this GLM poses a problem when attempting to make decisions about the better homerun hitter.

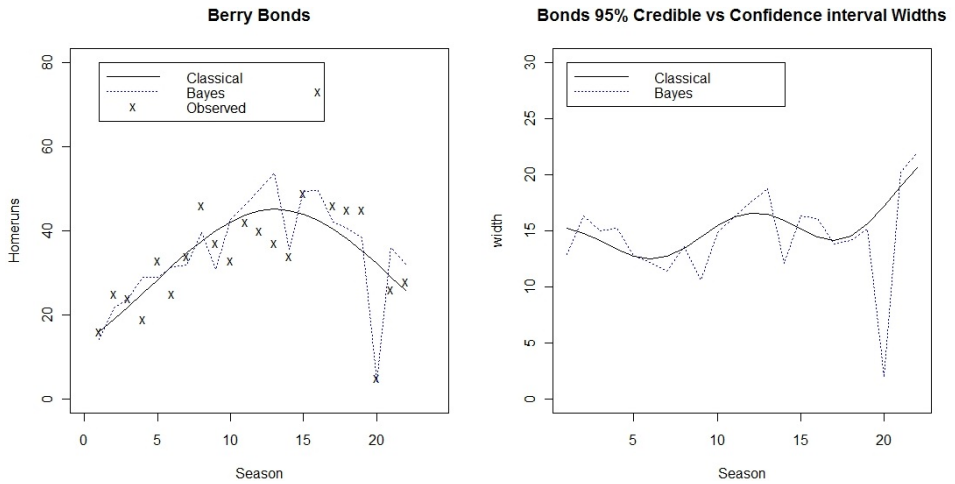

Figure 2.9: Baseball: Comparison of classical method and Bayesian method for Bonds.

If the interest was about *who hit the most homeruns* then clearly we would simply look at the observed counts and draw conclusions based on who had the most. However, our interest in is *who had the better homerun rate* suggesting that this model does not adequately answer our question of interest. The number of at bats for Sammy Sosa during these years is 643, 625, 604, 577, and 556 whereas McGuire only had 423, 540, 509, 521, 236, 299. With this information it is easy to see that the number of opportunities for Sosa to hit more homeruns would increase his overall homerun score, increasing the estimated expected number of homeruns which is what this GLM models. Clearly for this data this GLM is not appropriate for our question.

2.3.2 Incorporating At Bats

Using PROC GENMOD in SAS 9.2 with the Poisson distribution and log-link we analyzed the number of homeruns directly (using the log link) as a function of season, the square of season, and the number of at bats in the given season. Due to lack of fit we set the scale parameter to be estimated using the square root of the Pearson's Chi-Square statistic divided by it's degrees of freedom which is used

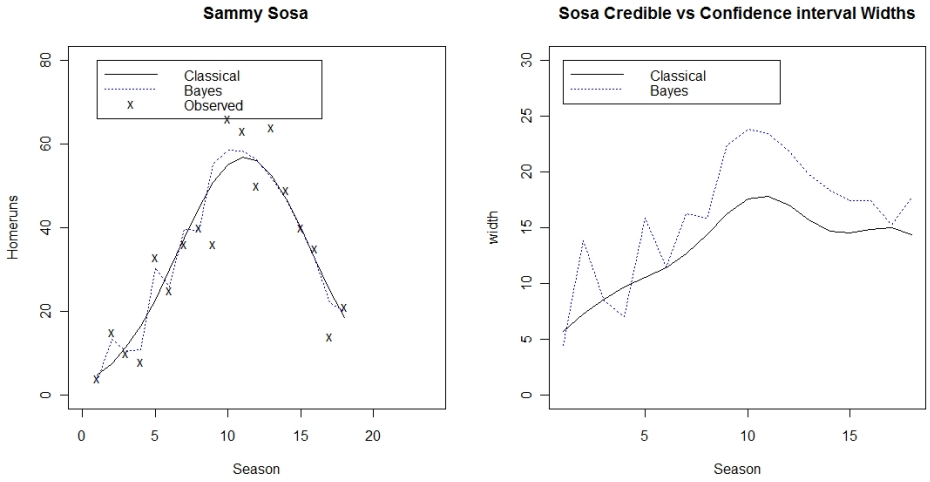

Figure 2.10: Baseball: Comparison of classical method and Bayesian method for Sosa.

to accommodate for overdispersion. As for the estimates of the coefficients we find similar results as found using model (2.1) and the previously considered classical model from section 2.3.1.

Here we display the estimated classical curves, the observed data, the estimated number of homeruns for each player as obtained by model (2.1), and a comparison of the 95% confidence intervals and 95% credible set widths.

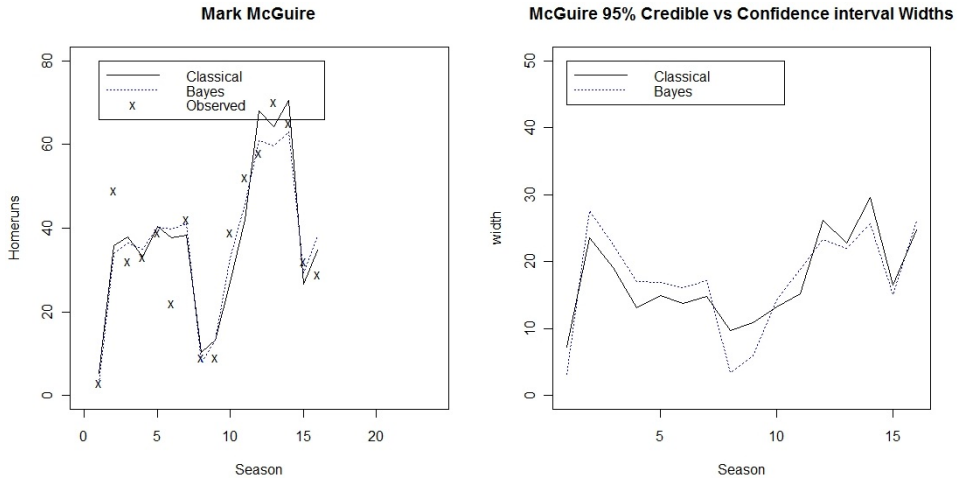

Figure 2.11: Baseball: Comparison of classical method, using at-bats, and Bayesian method for McGuire.

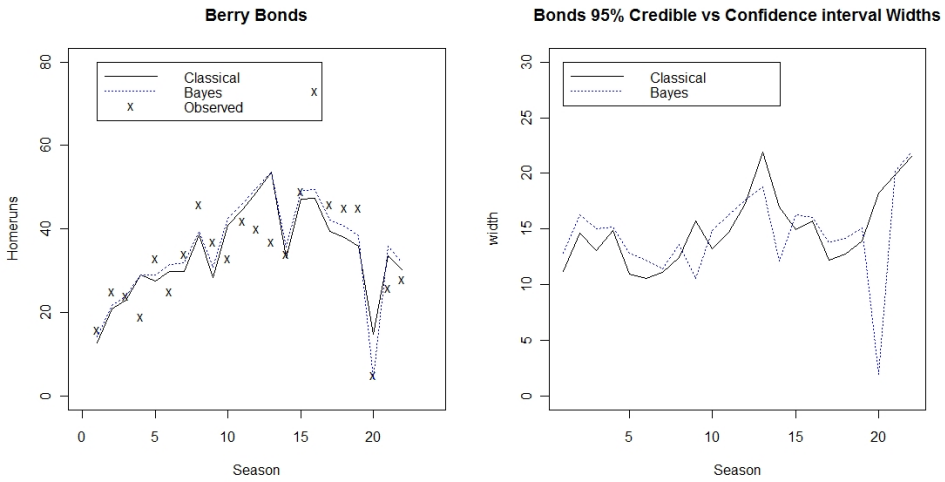

Figure 2.12: Baseball: Comparison of classical method, using at-bats, and Bayesian method for Bonds.

We notice in this case that the classical method and model (2.1) behave very similarly. Furthermore, incorporating the number of at bats for a player in a given season does a significantly better job in tracking the data than did the model from section 2.3.1. However, the AIC values for the model from section 2.3.1 were lower than those for the model incorporating the number of at bats leading a naive analyst to to conclude the model presented in section 2.3.1 to be a better choice.

It is worthy to note that the unscaled deviance, a measure of goodness of fit, for the GLM including the number of at bats for each player is significantly less than for the GLM which ignores this covariate; although the AIC is smaller for the later model. Such considerations need to be made rather than blindly following an AIC value when choosing an appropriate model.

We also make note that the credible set widths in the preceding analysis tend to be comparable to the 95% confidence interval widths. These credible set widths could be reduced arbitrarily depending on the chosen priors but caution must be taken as to not make such priors too informative. Since we used uniform flat priors over a relatively wide support for the given problem our intervals are rather wide. Furthermore, we could have used a few other powerhouse hitters of the day to

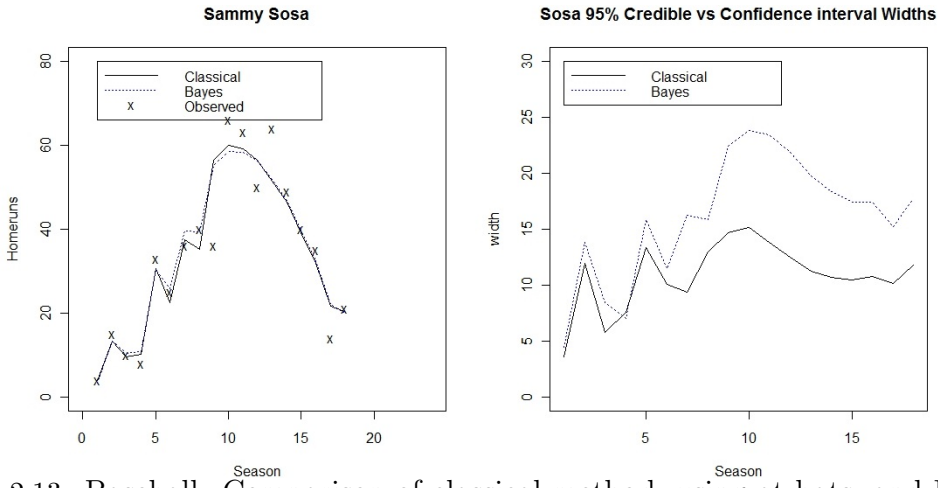

Figure 2.13: Baseball: Comparison of classical method, using at-bats, and Bayesian method for Sosa.

construct priors resulting in less posterior variance and thereby reducing the observed credible set widths.

2.4 Discussion

We've provided an example showing that the model developed in section 1.6 performed well for the data set at hand. Secondly, the construction of model (2.1) allows for the observed Poisson counts to be more closely fitted without *over smoothing* the data because the counts themselves are not directly modeled. In addition, model (2.1) is easily adaptable to any type of Poisson regression problem where overdispersion (underdispersion) in the data may be a concern through the use of the chi-square parameter.

Comparisons made with the classical counterpart showed that model (2.1) behaves similarly with wide flat priors. Furthermore, model (2.1) holds great potential in outperforming the classical approach giving more precise estimators provided good *a priori* information exists.

CHAPTER THREE

Habitat Burglary

A common definition of disease found on many websites, such as Merriam-Websters Online Dictionary, and hard print dictionaries, if not exact can be concisely worded as, *a condition of the living animal or plant body or of one of its parts that impairs normal functioning and is typically manifested by distinguishing signs and symptoms*. If viewed at a very simple level, crime is nothing more than an event within a society which disrupts the natural behavior of the victims involved. When related to the above definition of disease we may consider crime to be a pathogen and society or a community as the body within which the pathogen causes to malfunction or display symptoms.

Billions of dollars are spent each year by corporations, small businesses, and individuals on security personnel, security devices, and surveillance equipment in an effort to keep inventory and buildings secure. In cases where such measures are not adequate, thieves manage to pull off a heist. Lost inventory and labor are among the costs to the company/individual in replacing the stolen goods while also repairing any damage the thieves did to the facilities. Furthermore, the tax payer is now faced with the burden of police investigations and court costs. These are just a few of the observed symptoms of burglaries alone from the economic standpoint. With these considerations it is not unrealistic to consider crime a disease and use appropriate statistical modeling in an attempt to understand certain trends.

In this section we provide a spatial disease mapping example using habitat burglary data from the city of Waco Texas. We then combine the the model with an optimal subset selection procedure developed by Bratcher and Bhalla (1974) and used by Hamilton et al. (2008) and Stamey et al. (2004) to determine which

police beats have an elevated risk of habitat burglary. Furthermore, we consider the possibility that not all habitat burglaries were reported to the authorities and thus we attempt to correct for the naive estimates resulting in the analysis which does not allow this possibility.

3.1 The Methodology

Using a simplified version of model (1.15) which ignores any issues with under and over counts for this data set, let y_i be the observed count for region i and assume the following hierarchical structure,

$$\begin{aligned} Y_i|\lambda_i &\sim Poisson(t_i\lambda_i) \\ \lambda_i|(\mu_i,\alpha_i) &\sim Gamma(\mu_i\alpha_i,\alpha_i) \\ \alpha_i|\nu &\sim \chi^2(\nu) \\ log(\mu_i) &= \mathbf{x}\boldsymbol{\beta} + \epsilon_i \\ \boldsymbol{\beta} &\sim h(\cdot), \end{aligned} \tag{3.1}$$

where t_i is the expected number of events, λ_i is the local relative risk for region i, $h(\cdot)$ is a probability distribution, ϵ_i is a conditionally auto-regressive spatial error term, and ν is fixed. Notice that this means λ_i is a multiplicative factor able to adjust the local rate away from the expected t_i, which we will call the *inflation factor* or *relative risk*. Secondly, note that the mean of the gamma distribution is μ_i which results in a log-linear model on the mean of the inflation factors λ_i (local relative risk) rather than the Poisson rates themselves. We will calculate the t_i via internal standardization, discussed in section 1.4 under the hypothesis that the entire region has an overall constant burglary rate. This is done as,

$$t_i = n_i\overline{r} = n_i\frac{\sum y_i}{\sum n_i},$$

where n_i is the number of households at risk for region i and $\overline{r}$ may be thought of as an overall burglary risk.

An advantage of model (3.1) is that we model the mean of the inflation factors (relative risks) which allows for the data to be over/under dispersed naturally. Therefore, model (3.1) enables us to isolate regions with elevated local risks, associated with high inflation factors, as compared to the expected risk which differs from finding regions with the highest rate of occurrence thereby answering a different research question. Through the use of the ν parameter, we can adjust the sensitivity of the obtained model to the dispersion of the data; small values of ν will increase the variance of the Gamma distributed λ's. This can be observed by noticing that the variance of the gamma distribution, using the pdf described in equation (1.14), is $Var(\lambda_i|\alpha_i) = \frac{\mu_i}{\alpha_i}$. As this expression shows α, being dependent upon ν, can have considerable influence on the posterior distribution of λ_i. We comment on this further throughout this section.

3.2 Waco Crime Data

The data set that we analyze here was provided to us by the Waco Police department. It contains the number of received phone calls concerning habitat burglaries, broken down by police beats, for the City of Waco Texas during the year 2000. We use three covariates in our log-linear model. The first being the number of households below the poverty level, the second being the number of households that rent their habitat, and third is the area (in square miles) of the region in question. Covariate information is from the US Census Bureau's web site for the year 2000 and the area covariate was calculated using the boundary files and programs from R.

Since the Census data are given in census tracts or census blocks which do not correspond to the police beat regions provided to the authors, we used the program ArcGIS to dissolve the data to the appropriate police beat. The data is dissolved by considering the percent of area covered by a police beat from each

census tract then taking that percentage of the covariate values from that census tract, and repeating until each census tract a police beat covered is dissolved. We then aggregate each of these values together to obtain the values for the covariates for the police beats. Expected counts are calculated using internal standardization as discussed previously.

3.2.1 The Analysis

Figure 3.1 is a map of the observed standard mortality ratios (SMRs) by police beat. Notice the high density in the central part of the community. These regions are the police beats 8, 9, and 10, respectively.

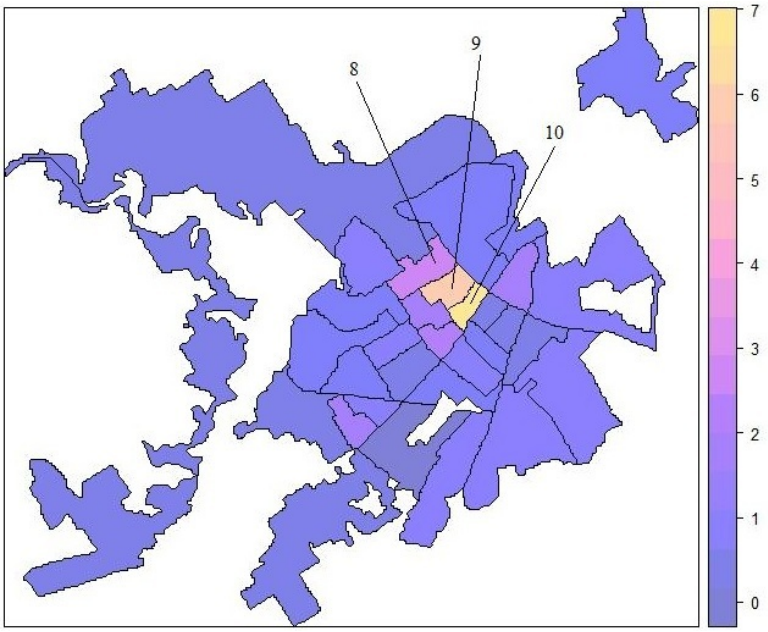

Figure 3.1: Raw SMR's by PD Beat

We point out that Figure 3.1 plots the SMRs not the raw burglary counts. Figure 3.2 presents the actual raw burglary counts. Notice the significant difference in the regions with high counts as opposed to high SMRs. If interest is which regions have the highest counts, we would focus on regions 14 and 22.

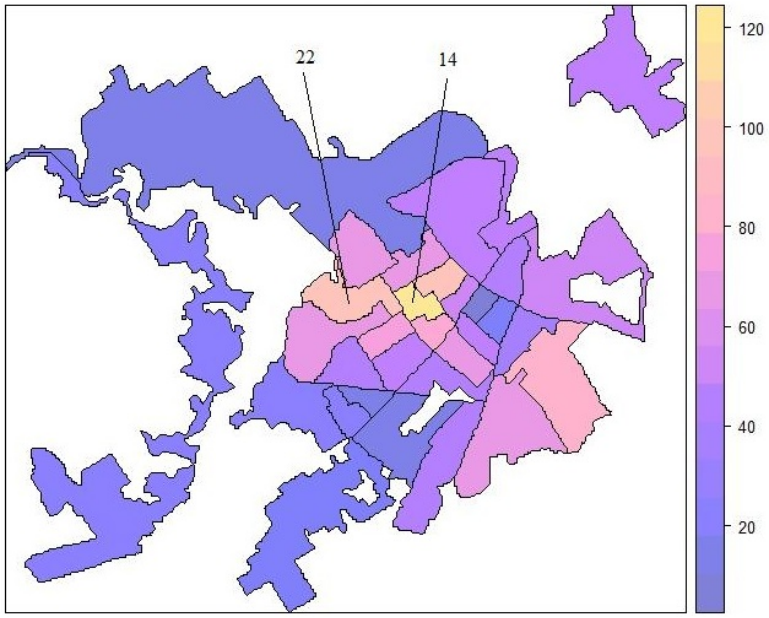

Figure 3.2: Raw SMR's by PD Beat

From this point on our focus will be on regions with high elevated risk as opposed to regions with the highest counts. This is done by dividing the burglary count by the expected count as calculated using the internal standardization method discussed in section 1.4.

We use model (3.1) to obtain estimates of the burglary rates of each region using an adjacency matrix where two regions are considered connected provided they share a common boundary. This adjacency matrix is then used in conjunction with the conditionally auto-regressive error term, ϵ_i, placed in the log-linear portion of the model. We ran the model with the alpha parameter having degrees of freedom of 5 and then applied the subset selection procedure (Bratcher and Bhalla, 1974) to the regions with the highest elevated risks.

Using model (3.1), we placed diffuse Normal$(0, 100)$ priors on the intercept, poverty, renter, and area terms where the notation used is mean and standard deviation. We place a $\chi^2(5)$ prior on the α_i terms. We used the program WinBUGS to find the posterior estimates provided from three chains with a 10,000 burin in and

Table 3.1: Posterior Estimates with $\alpha_i \sim \chi^2(5)$.

Covar.	Mean	MCErr.	Std.Dev	2.5%	Median	97.5%
Int.	0.2989	0.1377	7.56E-04	0.02546	0.2995	0.5668
Renter	-0.7843	0.2031	0.001486	-1.191	-0.7816	-0.3905
Poverty	0.4291	0.2085	0.001516	0.02742	0.4265	0.8503
Area	-0.08903	0.03713	1.98E-04	-0.1676	-0.08714	-0.02082

a sample size of 30,000 from each chain for a total of 90,000 after using a thinner of 30 to reduce autocorrelation. Convergence diagnostics were used to determine chain convergence and no issues were found: Gelman Rubin statistic, posterior density plots, and trace plots. The resulting posteriors for the beta terms were symmetric and bell shaped and the chains showed good mixing. Table 3.1 provides point estimates for the model where we have reported the estimated mean, standard deviation, Monte Carlo (MC) error, and the standard quantiles of interest. The Deviance Information Criterion (DIC) for this model was estimated to be 227.916.

Using the optimal subset selection procedure outlined in (Bratcher and Bhalla, 1974), with $c = 30$, we found that for all 30 regions only regions 9 and 10 had any nonzero probability associated with being the highest risk regions. These probabilities were 0.4132 and 0.5868 respectively. We then conclude that the regions 9 and 10 have the highest habitat burglary risk as compared to the rest of the community.

Figure 3.3 provides the spatial plot for the fitted inflation factors (relative risks) by police beat. Notice the high concentration in the center of the community associated with regions 8, 9, and 10. The interpretation of these results is that these regions are associated with higher risk for habitat burglary relative to their respective expected rates under the assumption of constant risk estimated via internal standardization. This is in agreement with the results of the optimal subset selection where regions 9 and 10 were found to be the most probable regions with the highest inflation factors. Lastly, the resulting residuals from model (3.1) are presented in Figure 3.4. Figure 3.5 displays the residuals from model (3.1) by predicted values.

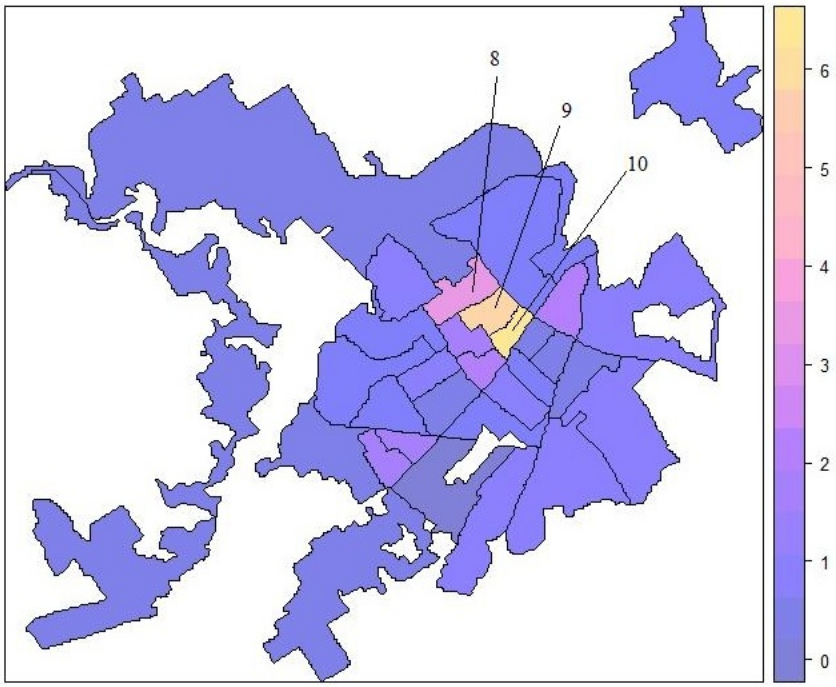

Figure 3.3: Fitted Inflation Factors for Bayes Model.

Recall that the parametrization of the Gamma distribution, as shown in equation (1.14), dictates the conditional variance, $Var(\lambda_i|\alpha_i) = \frac{\mu_i}{\alpha_i}$. This parametrization has the interpretation that for large values of α_i, more likely generated by a large ν value, we are very confident that the expectation of the inflation factors follow the specified log-linear behavior. When our confidence in this model is reduced, either by knowingly not including potentially significant covariates or due to over-dispersion of the data, we would reduce the ν parameter and allow for the Gamma variability to increase. We discuss this further in the next section.

3.2.2 Sensitivity Analysis

As reported by McBride 2006, and discussed briefly above, the posterior estimates are sensitive to the choice of the degrees of freedom parameter ν. In the analysis of section 3.2.1 we chose to use $\nu = 5$ to allow for moderate variability in the gamma distribution. Here we reanalyze the data using varying choices for the ν parameter leaving the same diffuse Normal(0,100) priors on the β's. We present the

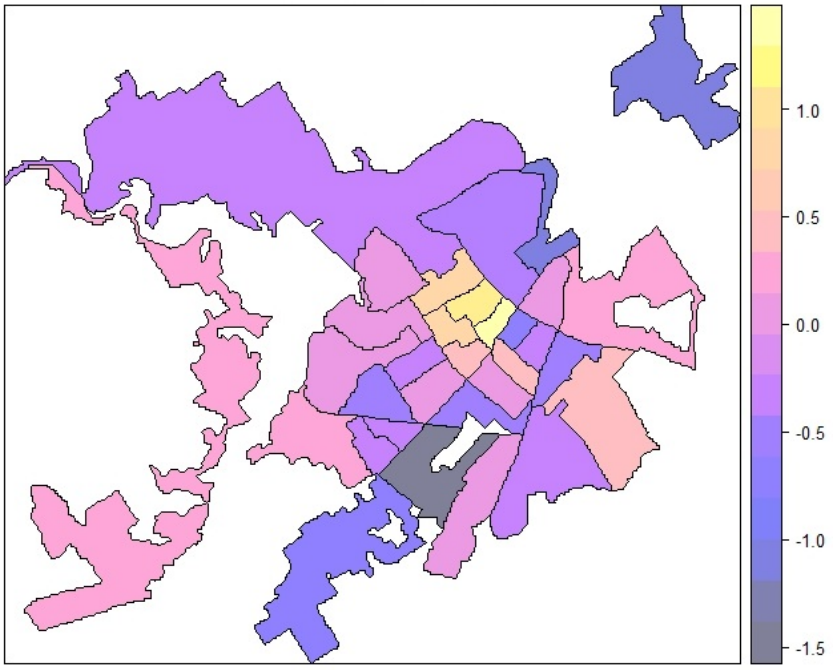

Figure 3.4: Residuals for Bayesian Model.

posterior mean and standard deviation for each of the covariate parameters in Table 3.2. The posterior mean estimates and standard deviations appear to stabilize near a ν value of 5 to 15.

Table 3.2: Posterior Estimates by choice of ν. *Thinner of 50 was used.

ν	Int. (SD)	Renter (SD)	Poverty (SD)	Area (SD)	DIC
0.01	0.84 (0.30)	-0.47 (0.42)	0.20 (0.45)	-0.040 (0.084)	229.197
0.1	0.75 (0.27)	-0.51 (0.38)	0.23 (0.40)	-0.052 (0.070)	229.196
0.5	0.75 (0.27)	-0.51 (0.38)	0.23 (0.40)	-0.052 (0.070)	229.196
1.0	0.43 (0.19)	-0.67 (0.26)	0.34 (0.27)	-0.079 (0.070)	228.572
2.0	0.35 (0.16)	-0.73 (0.23)	0.38 (0.24)	-0.086 (0.043)	228.034
5	0.30 (0.14)	-0.78 (0.20)	0.43 (0.21)	-0.089 (0.037)	227.897
10	0.29 (0.13)	-0.78 (0.19)	0.45 (0.20)	-0.089 (0.033)	227.932
15	0.28 (0.12)	-0.78 (0.19)	0.47 (0.20)	-0.089 (0.031)	228.045
20	0.28 (0.12)	-0.77 (0.19)	0.47 (0.20)	-0.089 (0.031)	228.173
25*	0.28 (0.12)	-0.77 (0.19)	0.48 (0.20)	-0.089 (0.031)	228.246

The autocorrelation appears to increase as the degrees of freedom for the χ^2 distribution increases. At $df = 20$ the auto correlation drops to zero by lag 25

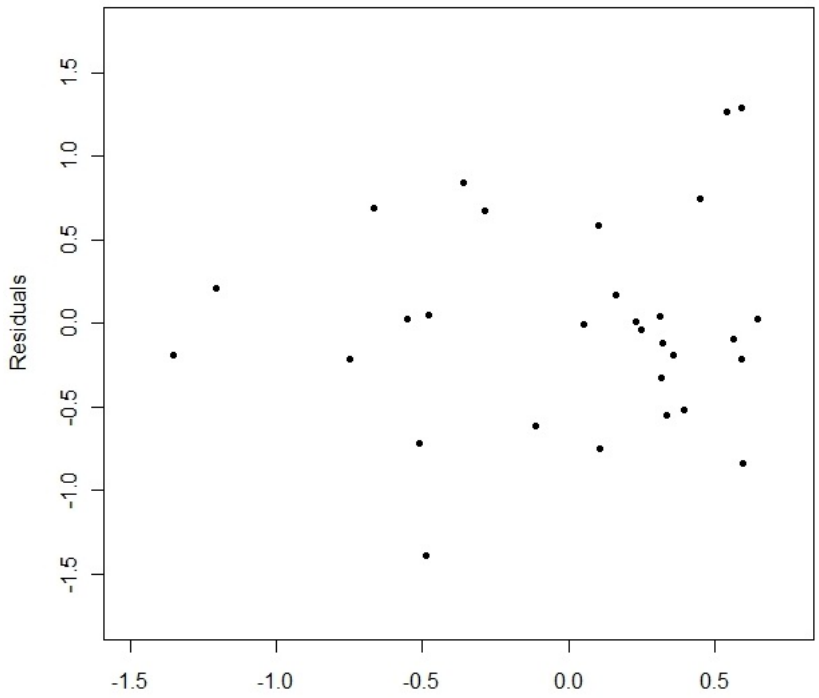

Figure 3.5: Bayesian Residuals by Predicted.

whereas a thinner of 50 is required to achieve the same goal when $df = 25$. The general trend observed in Table 3.2 is that as we put more *confidence* in the log-linear model the posterior standard deviation is reduced. As such, care must be taken when choosing the degrees of freedom parameter ν to allow for over-dispersion in the data if such behavior is either known *a priori* or suspected. The choice for ν does not seem to affects the posterior estimates for the PIC's obtained for the subset selection. Table 3.3 displays this effect for the regions 9 and 10 which shows that the PIC's remain fairly stable with differing choices of $\nu > 5$. We note that under all choices for ν only regions 9 and 10 showed non-zero probability for the PIC.

As the value for ν is decreased the difference between the PICs of regions 9 and 10 appears to get larger. This makes sense when taken into account for the fact that as ν goes to zero we are allowing the data to *speak* more loudly; our confidence in the log-linear relationship is decreased. For larger values of ν we are restricting the posterior distribution's variance to be quite small and about the expectation μ_i

which is a linear function of the predictors. The more we restrict the variability around this function the more the ordering of the μ_i's will affect the ordering of the PIC's. As such, similar covariates will generate similar expectations and the PICs will grow together.

Table 3.3: PIC for regions 9 and 10 by ν. *Thinner of 50 was used.

ν	Region 9	Region 10
0.01	0.3738	0.6262
0.1	0.3457	0.6543
1.0	0.3738	0.6262
2.0	0.3965	0.6035
5	0.4132	0.5868
10	0.4121	0.5879
15	0.4105	0.5895
20	0.4091	0.5909
25*	0.4101	0.5896

3.3 Under-reporting and Waco Habitat Burglary

It is a well known phenomenon that habitat burglary, like recreational use of marijuana, often goes unreported (Liska, Sanchirico, and Reed (1988); Academies (2005)). The factors influencing the victim(s) of habitat burglary to fail to report such crimes may stem from lack of confidence in the police to recover the lost items and apprehend the criminals responsible. Furthermore, the victim may have known the assailant(s) and due to this connection, possibly a familial one, wants to protect the individual(s). It is also possible that if such an investigation were to occur the police may find that the victim is involved in criminal behavior and so the loss of the stolen items is simply a price willing to be paid for their freedom and privacy.

3.3.1 The Analysis

Let us assume that the actual number of reported burglaries is less than actual number of habitat burglaries. Let us place a Beta(15.0342,2.5594) prior on the η

term in model (1.15). This beta distribution has 95% probability greater than 0.70 with a mode at 0.90. These values were chosen for the sake of demonstration rather than based on actual data that the reporting rates in Waco Texas would result in such a chosen prior. The resulting model becomes,

$$Y_i|\lambda_i \sim Poisson(t_i\lambda_i\eta), \tag{3.2}$$

$$\eta \sim beta(15.0342, 2.5594),$$

$$\lambda_i|(\mu_i, \alpha_i) \sim Gamma(\mu_i\alpha_i, \alpha_i),$$

$$\alpha_i|\nu \sim \chi^2(\nu),$$

$$log(\mu_i) = \mathbf{x}\boldsymbol{\beta},$$

$$\boldsymbol{\beta} \sim h(\cdot),$$

where t_i is the expected number of events, λ_i is the local relative risk for region i, $h(\cdot)$ is a probability distribution, η is the under-count report correction, and $\nu = 5$ is fixed. We choose for each β_i a Normal$(0, 10)$ prior where the notation used is mean and standard deviation. We used MCMC methods to obtain estimates for the posterior distribution. Three chains with a burnin of 30,000 iterations were used and after the burnin period a sample size of 90,000 was collected (three chains with 30,000 updates each) resulting in the estimates displayed in Table 3.4. A thinner of 30 was necessary to reduce auto correlation to zero by lag 20. Various convergence diagnostics were checked: trace plots, Gelman Rubin statistic, and density plots. Posterior densities were smooth and bell shaped. Deviance Information Criterion (DIC), originally proposed by Spiegelhalter, Best, Carlin, and van der Linde 2002 as a model selection tool, was estimated at 227.487.

The regions identified as having the highest habitat burglary risk were again regions 9 and 10 with probability of inclusions 0.411 and 0.589 respectively.

Table 3.4: Posterior Estimates for Under-Count model with Normal(0,10) priors.

Covar.	mean	sd	MC error	2.50%	median	97.50%
Int.	0.5201	0.1889	0.001041	0.16	0.5149	0.9092
Area	-0.08751	0.04044	1.80E-04	-0.1735	-0.08556	-0.01333
Poverty	0.3991	0.2219	0.001418	-0.03141	0.397	0.8408
Renter	-0.7637	0.2183	0.001379	-1.199	-0.7609	-0.3408

3.3.2 Sensitivity Analysis

As we are often concerned with posterior sensitivity to prior selection we rerun the analysis using more diffuse normal priors, Normal$(0, 31.62)$, with the same number of updates, thinning rate, and using a $\chi^2(5)$ distribution for the α parameter. The DIC was estimated at 227.686 and the posterior estimates for this model are displayed in Table 3.5. Notice that the results are quite similar as the results found in Table 3.4. The regions identified as having the highest habitat burglary risk were again regions 9 and 10 with probability of inclusions 0.4024 and 0.5976 respectively.

Table 3.5: Posterior Estimates for Under-Count model with Normal(0,31.62) priors.

Covar.	mean	sd	MC error	2.50%	median	97.50%
Int.	0.5189	0.1892	0.001163	0.1597	0.5136	0.906
Area	-0.08743	0.04052	1.99E-04	-0.1737	-0.08534	-0.01316
Poverty	0.4002	0.2237	0.001368	-0.03179	0.398	0.849
Renter	-0.7652	0.2193	0.00129	-1.205	-0.762	-0.3431

We also ran the model with very diffuse Normal$(0, 100)$ priors resulting in the posterior summary found in Table 3.6 and DIC of 227.658. The regions identified as having the highest habitat burglary risk were again regions 9 and 10 with probability of inclusions 0.4028 and 0.5971 respectively; the estimates are very similar.

Considering a more informative prior structure for the covariate terms we also ran the model with Normal$(0, 3.16)$ priors resulting in the posterior summary found in Table 3.7. The regions identified as having the highest habitat burglary risk were again regions 9 and 10 with probability of inclusions 0.4138 and 0.5832 respectively.

Table 3.6: Posterior Estimates for Under-Count model with Normal(0,100) priors.

Covar.	mean	sd	MC error	2.50%	median	97.50%
Int.	0.5208	0.1899	0.001124	0.1633	0.5161	0.9098
Area	-0.08744	0.04044	1.85E-04	-0.1733	-0.08538	-0.01404
Poverty	0.4003	0.2226	0.001338	-0.03223	0.3982	0.8456
Renter	-0.7652	0.2179	0.001293	-1.204	-0.7609	-0.3467

Table 3.7: Posterior Estimates for Under-Count model with Normal(0,3.16) priors.

Covar.	mean	sd	MC error	2.50%	median	97.50%
Int.	0.5164	0.1888	0.001184	0.1585	0.5107	0.9044
Area	-0.08699	0.04046	1.80E-04	-0.1728	-0.08498	-0.01252
Poverty	0.3939	0.2226	0.00142	-0.04094	0.3916	0.8365
Renter	-0.7596	0.2178	0.001391	-1.195	-0.757	-0.3383

Table 3.8 provides the widths of the obtained 95% credible sets. Notice that the widths are very similar.

3.4 Classical Comparison

In this section we perform the classical analysis of the same data, without accounting for under-reporting, using the comparable classical model structure for the purposes of comparison between classical analysis and the proposed model (3.1). That is, we use a log-linear relationship between the rates of the Poisson distributed burglaries and the covariates.

Using both a conditional auto-regressive (CAR) and a simultaneous auto-regressive (SAR) model, we model the habitat burglary log-rates as a function of

Table 3.8: 95% credible set widths.

SD (precision)	Int.	Renter	Poverty	Area
3.16 (0.1)	0.7459	0.16028	0.87744	0.8567
10 (0.01)	0.7492	0.16017	0.87221	0.8582
31.62 (0.001)	0.7463	0.16054	0.88079	0.8619
100 (0.0001)	0.7465	0.15926	0.87783	0.8573

the number of renter households, poverty, and region area in square miles. Both the renter and poverty variables are standardized. Although we only report the results for the SAR model, the analysis results are similar and the SAR model had a slightly lower Akakie-Information-Criterion (AIC) of 64.495 vs 66.949 for the CAR model.

Figure 3.6 plots the residuals for the SAR model by police beat. Figure 3.7 plots the residuals vs the predicted values for the SAR model. Figure 3.8 plots the fitted SAR model, the estimated burglary log-rates by region, for the community. The resulting model displays high risk of habitat burglary in the central part of the community relative to the regions on the outskirts.

Making comparisons between the Bayesian approach discussed in section 3.2.1, we notice similar results between the two approaches. Although model (3.1) and the classical approach place the log linear model on different outcomes the results are similar. That is, regions 8, 9, and 10 have the highest relative risk and Figures 3.3 and 3.8 display this. Figure 3.4 show very similar characteristics as the residuals from the classical approach displayed in Figure 3.7. As mentioned before model (3.1) places the log-linear relationship on a different response and so Figure 3.5 does not have the same scale for the predicted values as does Figure 3.7.

We make notice of the concentration of residuals in the middle of the community where the model under predicts the observed log-rates. These regions are 8, 9, 10, 13, and 14. We tested for spatial dependence failing to reject the null hypothesis of spatial correlation in these residuals (p-value = 0.11)

Parameter estimates and standard errors, obtained using the *spautolm* function found in the *spdep* package for R, are displayed in Table 3.9. The parameter estimates are similar to those found using model (3.1).

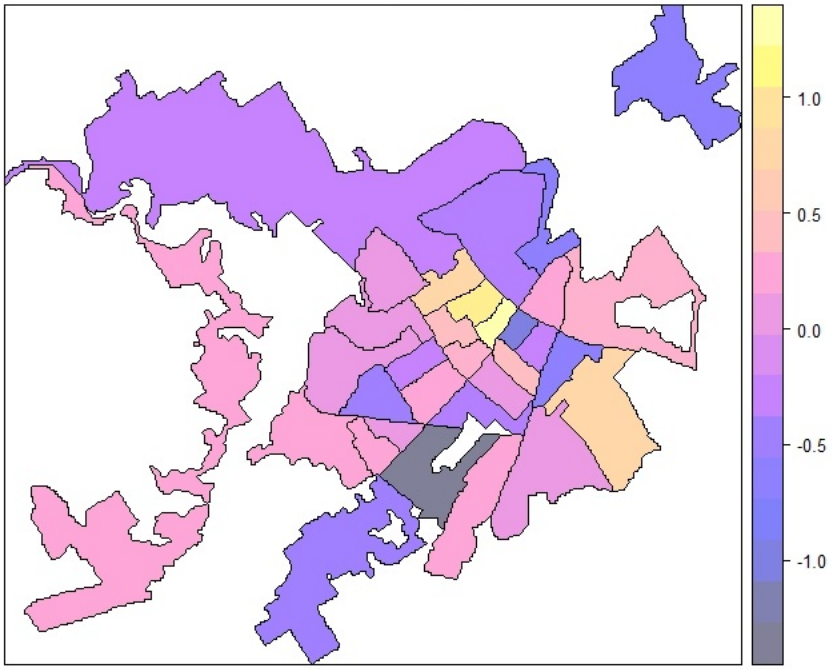

Figure 3.6: Residuals for SAR Model

Table 3.9: Parameter estimates and standard errors.

	Estimate	Std. Error	P-Value
Renter	-0.660	0.162	< 0.0001
Poverty	0.534	0.189	< 0.0001
Area	-0.079	0.0213	0.0002

3.5 Reparameterization of the Hierarchical Model

As we mentioned in the Introduction, there are several classical methods to account for overdispersion in the response. Some popular methods assume a mean-variance relationship and specify the polynomial degree of the relationship such as linear or quadratic; see section 1.1.2. For (3.1), this relationship is linear with the conditional variance $\frac{\mu_i}{\alpha_i}$. If the researcher believes that this relationship is quadratic the parameterization of (3.3) found below provides the expectation μ_i on the λ_i's and has conditional variance, $\frac{\mu_i^2}{\alpha_i}$,

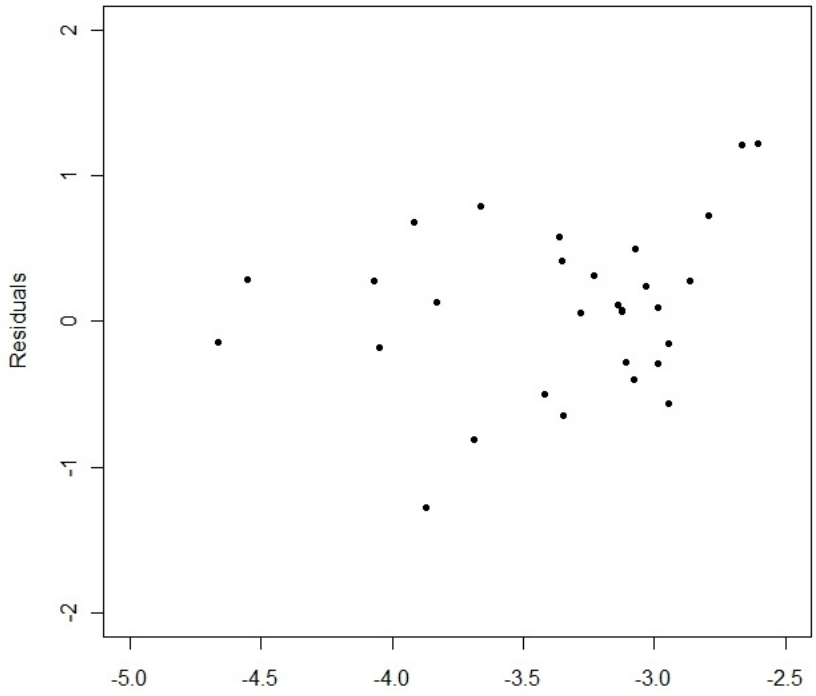

Figure 3.7: Predicted vs. Residuals for SAR Model

$$Y_i|\lambda_i \sim Poisson(t_i\lambda_i) \tag{3.3}$$

$$\lambda_i|(\mu_i, \alpha_i) \sim Gamma(\alpha_i, \frac{\alpha_i}{\mu_i}),$$

$$log(\mu_i) = \mathbf{x}\boldsymbol{\beta} + \epsilon_i,$$

$$\boldsymbol{\beta} \sim h(\cdot),$$

where all variables represent the same values as model (3.1).

We reanalyze the Waco habitat burglary data set, ignoring the issue of under counts here, using the parameterization of (3.3) where $\alpha_i \sim \chi^2(5)$ and Normal$(0, 100)$ priors for the β terms. Three chains with a burnin of 10,000 iterations was used followed by a 30,000 iterations from each chain for the sample giving a total sample size of 90,000. A thinner of 30 was used to reduce autocorrelation. Table 3.10 provides the posterior estimates for the mean, standard deviation, and the 95% credible sets. Although the values are different than those obtained using model (3.1), we do notice that the estimates are roughly similar. Regions 9 and 10 are

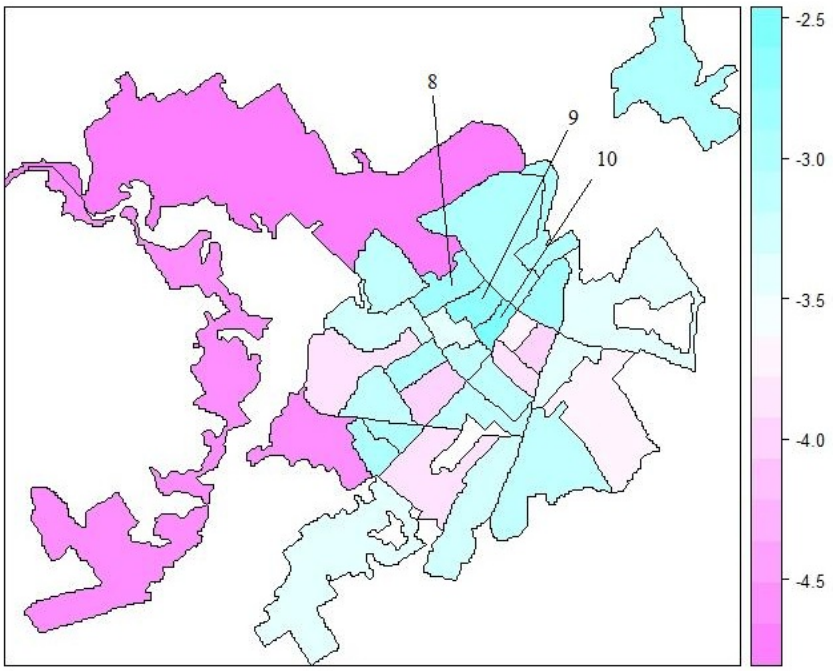

Figure 3.8: Fitted log-rates for SAR Model.

Table 3.10: Posterior Estimates using model (3.3).

Covar.	mean	sd	MC error	2.50%	median	97.50%
Int.	0.3328	0.1351	5.59E-04	0.06552	0.3325	0.5991
Renter	-0.7616	0.1986	0.001148	-1.159	-0.76	-0.3755
Poverty	0.4725	0.2103	0.001238	0.06743	0.4691	0.8937
Area	-0.08905	0.03014	1.33E-04	-0.1482	-0.0892	-0.02909

again the only regions with associated non-zero PICs; 0.3897 for region 9 and 0.6103 for region 10. Using the parameterization found in model (3.3), care is still required in choosing the value for ν as the variance for the gamma distributed λ_i's is related to the value of α_i.

3.6 Discussion

The proposed hierarchical Bayesian model (3.1) provides the researcher with the ability to control for overdispersion in their data with the $\alpha_i \sim \chi^2(\nu)$ parameter. However, caution must be taken when choosing the degrees of freedom associated

with this chi-square distributed parameter as autocorrelation in the MCMC chains and significant bias toward an equal mean-variance relationship can occur. This parameter affects the posterior distribution of the covariate parameters. However, for the Waco burglary data we found that the PICs the parameters tend to become stable for $\nu \geq 5$. This is promising if the research interest lies in determining regions with high associated risk but not parameter estimates and relationships between chosen covariates and the outcome variable.

Model (3.1) can be easily reparameterized to allow for a quadratic mean-variance relationship using model (3.3) demonstrating great flexibility in the variance-mean relationship. This variance flexibility is not limited to simple linear and quadratic behavior but can be constructed for any power of μ_i by adjusting the rate and shape parameters of the gamma distribution while retaining the conditional mean of μ_i. Both linear, model (3.1), and quadratic variance, model (3.3), relationships provide results that are similar for the data we analyzed here. Furthermore, model (3.1) provided estimates similar to those found using comparable classical methods. Combined with the addition of the subset selection procedure the researcher is able to easily provide inference, in terms of probability, on the regions associated with the highest relative risk and inference on the parameters chosen.

Model (3.1) differs from the usual log-linear model where we place the log-linear relationship on the mean of the Poisson rates rather than the Poisson rates themselves. This allows us to use the ν parameter to adjust our confidence in the Poisson assumption of equality between mean and variance. By using larger values of ν we place more confidence in this assumption *essentially* making the claim that over-dispersion in the data is small or non-existent.

Finally, the limitations of any model derived from (1.15), using independent priors for the covariate terms, is that for large values of ν the autocorrelation in the Markov chains becomes intolerable. This situation will occur in data sets where

large expected counts, E_i, are common but over-dispersion is believed to be minimal, say, variance is twice the mean. This limitation is a software based limitation from what the authors have noticed using WinBUGS. To date, no study has been done on this phenomenon using another MCMC software package.

CHAPTER FOUR

A New Model

In the previous chapters we have taken McBride's (2006) hierarchical model and modified it to account for under counts in the data. We reexamined the homerun and Waco habitat burglary examples he provided and compared them to a classical analysis finding similar results for an appropriate choice of the parameter ν. Unfortunately, what we discovered is that the model proposed by McBride, when using independent priors for the covariate terms, performs poorly in the software program WinBUGS; as autocorrelation for large values of ν, which correspond to moderate to high confidence in a Poisson assumption of equality between mean and variance for the system under study, becomes intolerable for complex data sets. This autocorrelation in the Monte-Carlo Markov chains calls into question any inference made and makes the proposed model, although theoretically reasonable, impractical for common applications. Furthermore, in order for under-dispersion to be accounted for the scale parameter $(1 + E_i\frac{1}{\nu-2})$ must be less than one. This requires that, $-1 \leq E_i\frac{1}{\nu-2} < 0$, in order to account for under-dispersion while retaining a positive variance.

What we propose here is a more flexible approach to model (1.15). Our proposed model is,

$$Y_i|\lambda_i \sim Poisson(E_i\lambda_i\eta) \tag{4.1}$$

$$\eta|(a,b) \sim Beta(a,b)$$

$$\lambda_i|(\mu_i,\alpha_i) \sim Gamma\left(\frac{\mu_i}{\alpha_i},\frac{1}{\alpha_i}\right)$$

$$\alpha_i|(c,d) \sim Beta(c,d)$$

$$\mu_i = \gamma_i\kappa$$

$$log(\gamma_i) = \mathbf{x}\boldsymbol{\beta}$$

$$\boldsymbol{\beta} \sim h(\cdot),$$

with the same stipulations as those in model (1.15). In model (4.1) we place a Beta prior on the α_i terms instead of a χ^2 and change the parameterization of the Gamma distributed λ_i's. The Beta distribution is rich in shapes allowing for the researcher great flexibility in their confidence of the Poisson assumption. Furthermore, the beta(c, d) distribution, for constants c and d, can be transformed from the default support of $(0, 1)$ to any finite support, (r, s), using the transformation, $z = (s - r)x + r$. The transformed Beta distribution will then have mean $(s - r)(\frac{c}{c+d}) + r$. Using model (4.1) provides the following variance for the observables Y_i,

$$\begin{aligned}
Var(Y_i|E_i) =& E(Var(Y_i|\lambda_i, E_i)) + Var(E(Y_i|\lambda_i, E_i)) \\
=& E(E_i\lambda_i) + Var(E_i\lambda_i) \\
=& E_i\mu_i + E_i^2 Var(\lambda_i) \\
=& E_i\mu_i + E_i^2(E(Var(\lambda_i|\alpha_i)) + Var(E(\lambda_i|\alpha_i))) \\
=& E_i\mu_i + E_i^2(E(\mu_i\alpha_i) + Var(\mu_i)) \\
=& E_i\mu_i + E_i^2(\mu_i E(\alpha_i) + 0) \\
=& E_i\mu_i + E_i^2\left(\mu_i\frac{c}{c+d}\right) \\
=& E_i\mu_i\left(1 + E_i\frac{c}{c+d}\right).
\end{aligned} \tag{4.2}$$

An advantage of model (4.1) over model (1.15) is that the high autocorrelation phenomenon when using WinBUGS seems to less problematic when values near zero are chosen for the mean of the $\alpha_i \sim beta(c, d)$ parameter. This corresponds to choosing large values of ν in model (1.15); that is, *a-priori* belief that over-dispersion is small. Furthermore, model (4.1) enables a correction for under-dispersion, allows for great flexibility in distribution shapes for the α_i parameter, and seems to perform better in WinBUGS for the data sets analyzed in this dissertation.

In the remaining portion of this chapter we reanalyze the homerun and habitat burglary data using model (4.1).

4.1 Habitat Burglary Revisited

We use the same covariates and model (4.1) we present the analysis of the Waco Habitat burglary here. We place diffuse Normal$(0, 100)$ priors on the intercept, poverty, renter, and area terms where the notation used is mean and standard deviation. We place a Beta$(2.74, 34.20)$ prior on the α_i terms. This beta has a mean of 0.074, median 0.067, mode 0.05, and variance 0.0018. Although this prior is informative we believe that for expected counts of 50, on average, we would see a variance inflation of about 4.5 and for expected counts of 100 we would see an inflation of 8. We have no prior data to base this expectation on other than the observed data. Therefore, this choice is for demonstration purposes. We will ignore the η parameter with the assumption that no reports go uncounted for the time being.

We used the program WinBUGS to find the posterior estimates provided from three chains with a 10,000 burin in and a sample size of 30,000 from each chain for a total of 90,000 after using a thinner of 20, which is less than the thinning rate used before to analyze the same data, to reduce autocorrelation. Convergence diagnostics were used to determine chain convergence and no issues were found: Gelman Rubin statistic, posterior density plots, and trace plots. The resulting posteriors for the beta terms were symmetric and bell shaped and the chains showed good mixing. Table 4.1 provides point estimates for the model where we have reported the estimated mean, standard deviation, Monte Carlo (MC) error, and the standard quantiles of interest. Notice the parameter estimates are similar to those found in section 3.2.2 for $\nu \geq 10$. The DIC for this model was estimated to be 227.902, which is very close to the DIC's found in section 3.2.2.

Table 4.1: Posterior estimates where $\alpha_i \sim Beta(2.74, 34.20)$.

Covar.	Mean	Std.Dev	MCErr.	2.5%	Median	97.5%
Intercept	0.2893	0.1289	9.43E-04	0.03189	0.2904	0.5397
Renter	-0.7833	0.1947	0.002095	-1.173	-0.7823	-0.3994
Poverty	0.4462	0.2008	0.002179	0.05504	0.4455	0.8477
Area	-0.08897	0.03398	2.33E-04	-0.159	-0.08814	-0.02442

Figure 4.1 provides the spatial plot for the fitted inflation factors (relative risks) by police beat. Notice the high concentration in the center of the community associated with regions 8, 9, and 10. The interpretation of these results is that these regions are associated with higher risk for habitat burglary relative to their respective expected counts under the assumption of constant risk estimated via internal standardization. This is in agreement with the results of the optimal subset selection where regions 9 and 10 were found to be the most probable regions with the highest inflation factors. Lastly, the resulting residuals from model (4.1) are presented in Figure 4.2. Figure 4.3 displays the residuals from model (4.1) by predicted values.

Using the optimal subset selection procedure outlined in (Bratcher and Bhalla, 1974), with $c = 30$, we found that for all 30 regions only regions 9 and 10 had any nonzero probability associated with being the highest risk regions. These probabilities were 0.4133 and 0.5867 respectively, nearly identical results previously reported in section 3.2.2 for larger values of ν. We then conclude that the either regions 9 and 10 have the highest habitat burglary risk as compared to the rest of the community.

We now concern ourselves with the possibility that not all burglaries are reported. What we provide here is an analysis similar to that found in section 3.3 where we use model (4.1). No matter the factors influencing the victims decision not to report let us assume that the actual number of reported burglaries is less than actual number of habitat burglaries. Let us place a Beta(15.0342,2.5594) prior on the η term in model (4.1). This beta distribution has 95% probability greater than

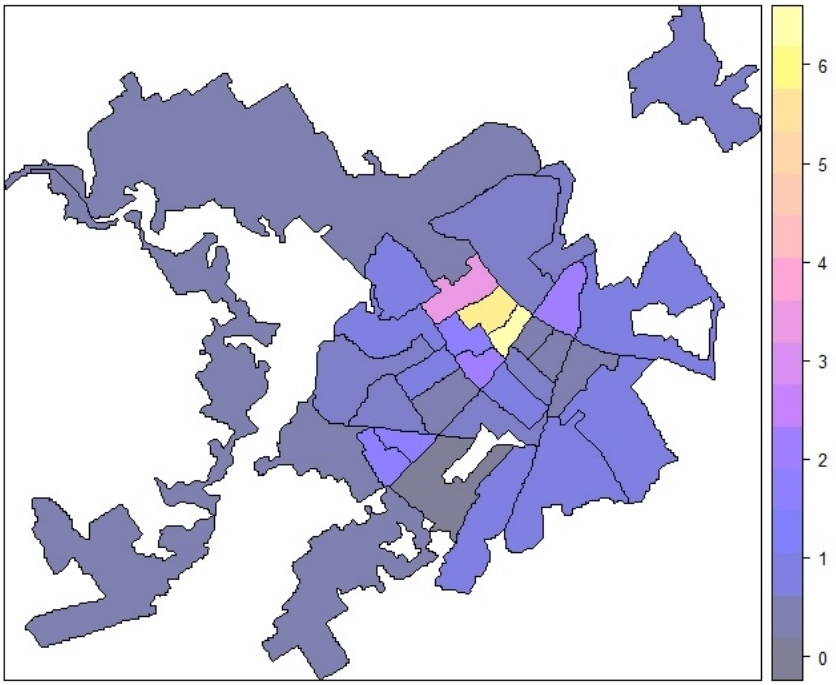

Figure 4.1: Fitted inflation factors for Bayes model (4.1).

0.70 with a mode at 0.90. These values were chosen for the sake of demonstration rather than based on actual data that the reporting rates in Waco Texas would result in such a chosen prior.

We used the program WinBUGS to find the posterior estimates provided from three chains with a 30,000 burin in and a sample size of 30,000 from each chain for a total of 90,000 after using a thinner of 22 to reduce autocorrelation; only a burning of 10,000 was needed and we simply chose to leave out the first 30,000 to keep similar sample sizes as used before. Furthermore, we note that this model had roughly the same thinning rate and converged after the same number of iterations of the chain. This was not the case with model (3.2) where a significantly higher thinning rate was needed; in the previous analysis we needed a thinner of 30 for $\nu = 15$.

It is possible that the observed decrease in thinning rate was due to a number of different reasons. These may include an interaction between the specific data available and the computer platform were were using, the specific chosen beta dis-

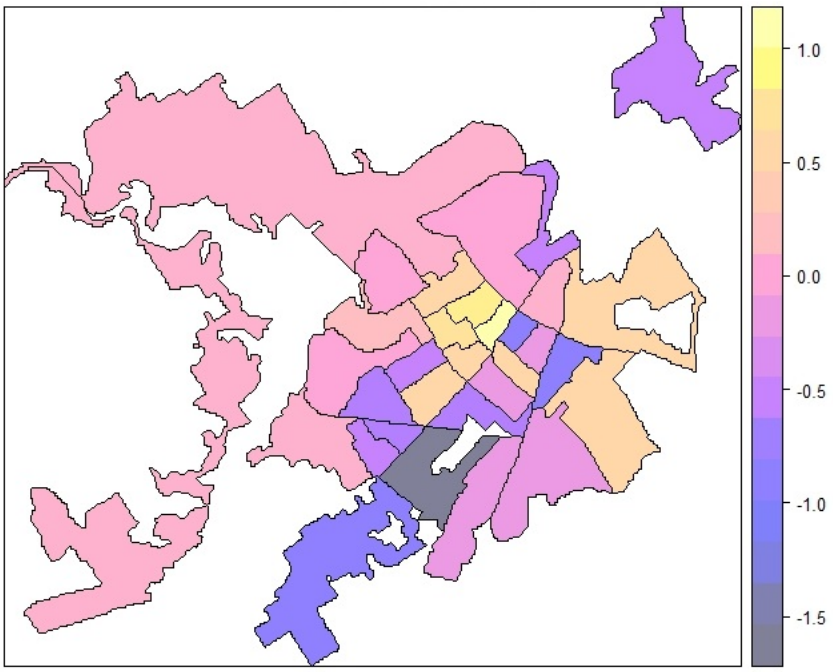

Figure 4.2: Residuals for Bayesian model for (4.1).

tribution, the limited support of the beta distribution, or several other factors. We seem to have stumbled on an example where we have better performance using our proposed model as opposed to (1.15). Therefore, we believe that the current parameterization seemed to have behaved better in this application and may show promise in out performing (1.15) in general but we have not tested this assertion.

Convergence diagnostics were used to determine chain convergence and no issues were found: Gelman Rubin statistic, posterior density plots, and trace plots. The resulting posteriors for the beta terms were symmetric and bell shaped and the chains showed good mixing. We make note that a thinner of only 22 was need to obtain acceptable autocorrelations, zero by lag 20, whereas for model (3.2) with large ν significantly larger thinning intervals was needed to accomplish the same goal. Table 4.2 displays the associated point estimates for the covariate parameters.

Using the optimal subset selection procedure, with $c = 30$, we found that for all 30 regions only regions 9 and 10 had any nonzero probability associated with

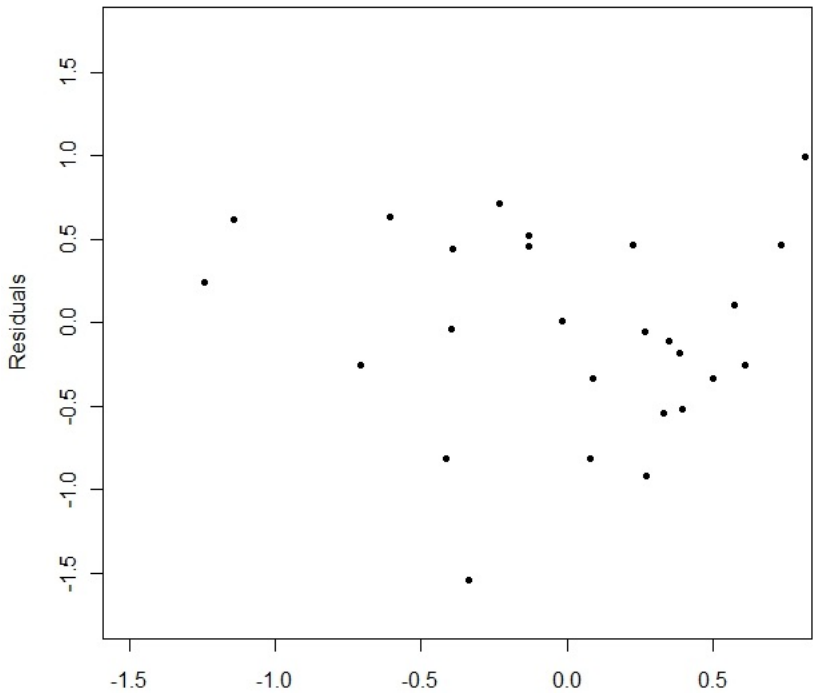

Figure 4.3: Bayesian residuals by predicted for (4.1).

being the highest risk regions. This is a reoccurring theme with this data set. These probabilities were 0.4108 and 0.5892 respectively.

Figure 4.4 provides the spatial plot for the fitted inflation factors (relative risks) by police beat. Notice the high concentration in the center of the community associated with regions 8, 9, and 10. The interpretation of these results is that these regions are associated with higher risk for habitat burglary relative to their respective expected rates under the assumption of constant risk estimated via internal

Table 4.2: Posterior estimates where $\alpha_i \sim Beta(2.74, 34.20)$ and $\eta \sim Beta(15.03, 2.56)$.

Covar.	Mean	Std.Dev	MCErr.	2.5%	Median	97.5%
Intercept	0.4538	0.1627	0.00119	0.1515	0.4472	0.7964
Renter	-0.7781	0.1916	0.001677	-1.16	-0.777	-0.4036
Poverty	0.4481	0.1989	0.00172	0.05919	0.4471	0.8445
Area	-0.08887	0.03343	1.85E-04	-0.1576	-0.08799	-0.0254

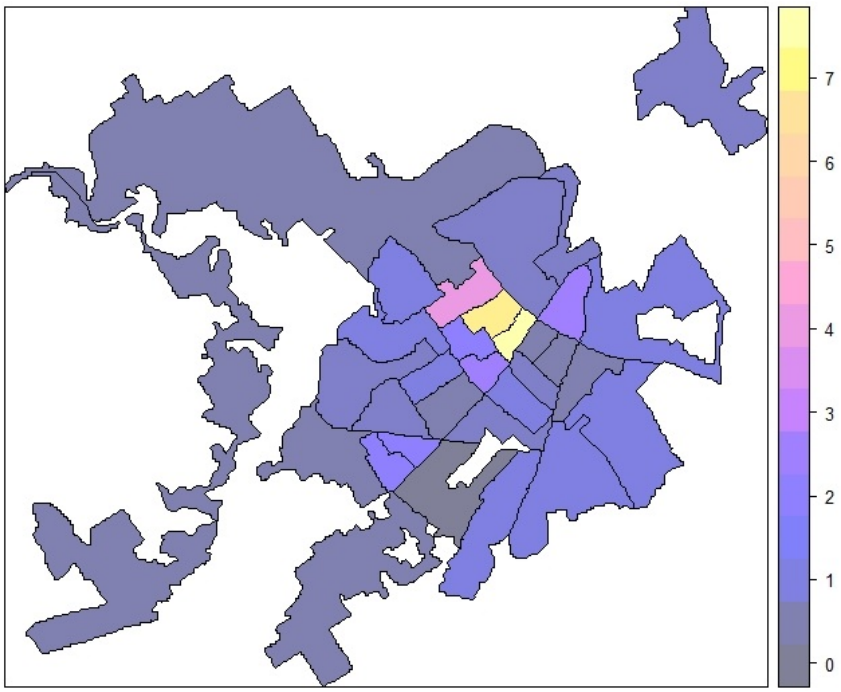

Figure 4.4: Fitted Inflation Factors for Bayes Model for model (4.1).

standardization. This is in agreement with the results of the optimal subset selection where regions 9 and 10 were found to be the most probable regions with the highest inflation factors. Lastly, the resulting residuals from model (4.1) are presented in Figure 4.5. Figure 4.6 displays the residuals from model (4.1) by predicted values.

4.2 Homerun Data Revisited

We reanalyze the homerun data from chapter 2 here using model (4.1) and ignore the possibility of undercounts as we are certain all homeruns were accounted for. The majority of at bats ranges from about 200 to just over 500 for each player. We do not expect to see a variance inflation, using equation (4.2), greater than 3 over the mean. That said, we choose to place an informative Beta(1.24, 48.60) prior on the α_i terms. This Beta distribution has a mean of 0.025, variance of 4.7E-4, mode of 0.005, with the lower 2.5% and upper 97.5% percentiles of 0.0012 and 0.082 respectively.

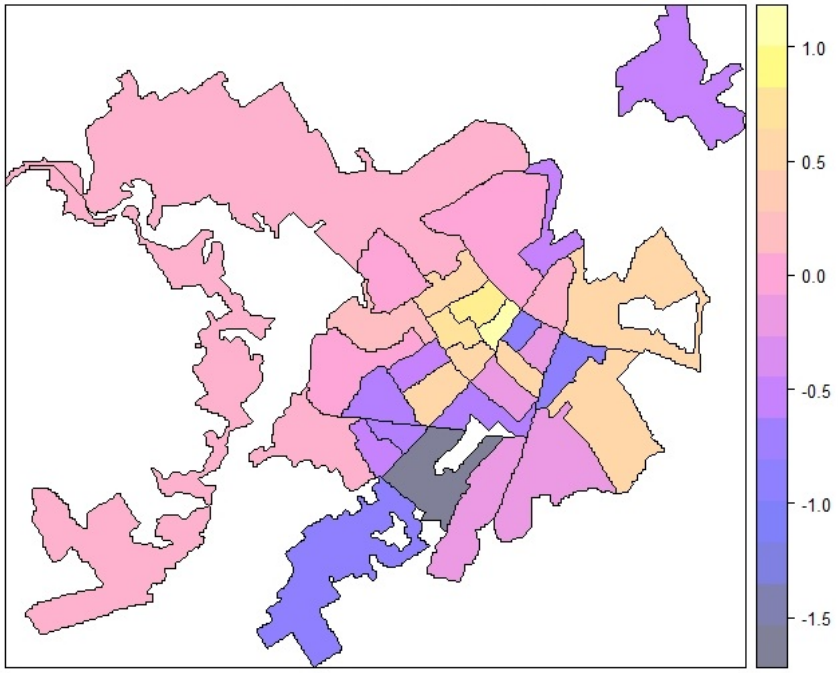

Figure 4.5: Residuals for Bayesian Model for model (4.1).

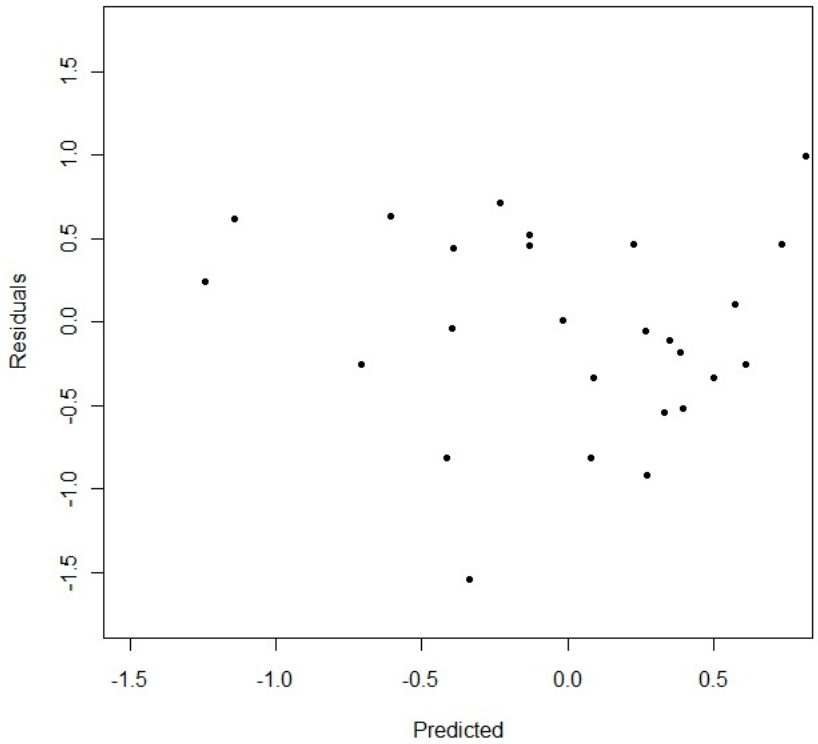

Figure 4.6: Bayesian Residuals by Predicted for model (4.1).

With the consideration of the expected behavior as discussed in section 2.2.1 we set uniform priors of Uniform(-10,0), Uniform(-3,5), Uniform(-4,4) for the terms β_0, β_1, and β_2, respectively. Notice that these priors are considerably wider than those chosen in section 2.2.2. We allow for some probability weight on positive values for the quadratic term in the event that a player does not follow a decreasing trend in his later seasons. We used three chains in WinBUGS with a burnin of 30000, thinner of 5, and further updates of 30,000 for a total of 90,000 posterior samples for our point estimates. Convergence diagnostics were used to determine chain convergence and no issues were found: Gelman-Rubin statistic, autocorrelation plots, trace plots and density plots. Posterior density estimates for the linear, quadratic, and intercept terms were symmetric and bell shaped. Table 4.3 presents the associated posterior point estimates.

Table 4.3: Parameter estimates for (4.1) where $\alpha_i \sim Beta(1.24, 48.60)$.

Player	Para.	Mean	SD	MC Err.	2.5%	Median	97.5%
McGuire	Int.	-2.303	0.153	8.0E-04	-2.618	-2.300	-2.014
	Lin.	0.054	0.024	9.5E-05	0.009	0.054	1.05E-01
	Quad.	-0.003	0.006	3.2E-05	-0.015	-0.002	9.40E-03
Bonds	Int.	-2.364	0.127	5.9E-04	-2.615	-2.364	-2.115
	Lin.	0.047	0.015	5.6E-05	0.017	0.046	7.77E-02
	Quad.	-0.004	0.003	1.3E-05	-0.009	-0.004	1.17E-03
Sosa	Int.	-2.388	0.140	6.3E-04	-2.672	-2.385	-2.12E+00
	Lin.	0.054	0.024	8.5E-05	0.008	0.0537	1.02E-01
	Quad.	-0.014	0.005	2.2E-05	-0.025	-0.0143	-4.58E-03

Notice that for McGuire, the 95% credible set for the quadric term includes zero, although the upper bound is near zero. We choose to leave the quadratic term in the model since this upper bound is near zero. The quadratic term for Bonds appears to be needed as the upper bound is practically zero. The 95% credible set for the quadric term for Sosa, as expected, excludes zero.

To answer the question of which season was the best season for each of the three players turn the the subset selection procedure. Our goal is to select which season was the peak season for the individual players and select from the players which had the highest homerun rate at the peak season. We use here the optimal subset selection procedure developed by Bratcher and Bhalla (1974) to answer the question at hand. A constant loss function is chosen with the ratio $c = 24$.

We present here the probabilities of each season being the best season for hitting homeruns by each player in table 4.4. Using the subset selection procedure discussed previously, with $c = 24$, we would conclude that McGuires best seasons were either of the seasons 8, 10, 11, 13, 14, or 15; Bond's best seasons were either of the seasons 16, 19, or 20; Sosa's best seasons were either of 10, 11, 13, or 14.

Table 4.4: PICs for each season using model (4.1).

Season	Bonds	McGuire	Sosa	Season	Bonds	McGuire	Sosa
1	0	0.004992	0	12	2.15E-05	0.01224	0.02056
2	0	1.44E-04	0	13	0	0.2911	0.4791
3	0	0	0	14	0.004811	0.1079	0.04058
4	0	0	0	15	0.006296	0.3178	0.006849
5	0	1.03E-05	1.04E-05	16	0.6987	0.01846	0.00358
6	0	0	2.48E-04	17	0.03436	NA	1.04E-05
7	0	6.68E-04	2.59E-04	18	0.03964	NA	2.07E-05
8	1.61E-04	0.06562	0.01367	19	0.06804	NA	NA
9	0.002013	0.003431	2.07E-05	20	0.1447	NA	NA
10	0	0.0941	0.2329	21	8.61E-05	NA	NA
11	9.69E-05	0.08351	0.2022	22	0.001087	NA	NA

To answer the question of which of the three players was the best homerun hitter over their career we also use the subset selection procedure. We answer the question, "Which of the three players had the highest career peak?" We performed the analysis in WinBUGS where we estimated the covariate terms for each of the three players simultaneously, independently of each other, and then calculated the PIC for each player having the highest career peak. The PIC for McGuire was esti-

mated at 0.4601, for Bonds the PIC estimate was 0.5363, and for Sosa the estimated PIC was 0.0036. Using our decision rule, we conclude that either Bonds or McGuire was the better homerun hitter of the three players. If interest was in which of either Bonds or McGuire was the better hitter, we estimate the probability that Mcguire had a better career max than Bonds at 0.4619.

Figure 4.7 displays the estimated player rates on the same graph for comparison. Figures 4.8, 4.9, and 4.10 display the estimated number of homeruns for both the classical analysis found in section 2.3.2 and model (4.1) for comparisons. We make notice of the fact that both methods tend to track the observed homeruns well with model (4.1) being slightly more accurate, although with wider probability intervals than the classical confidence limits.

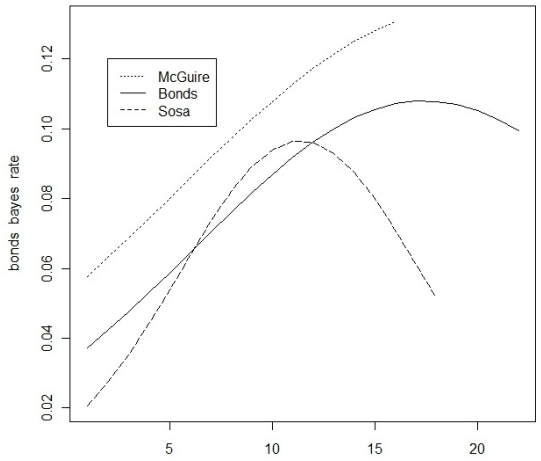

Figure 4.7: Comparison of players estimated curves using model (4.1).

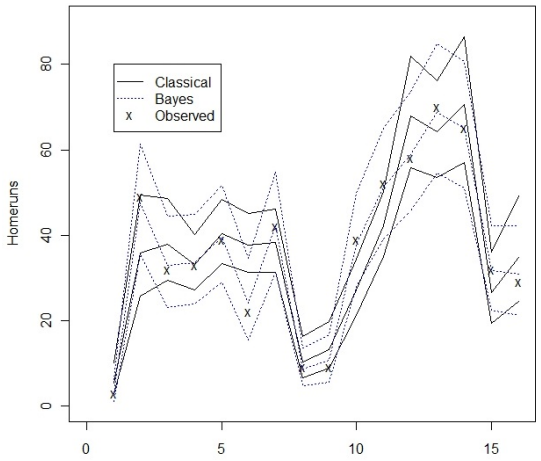

Figure 4.8: Comparison of classical method, using at-bats, and model (4.1) for McGuire.

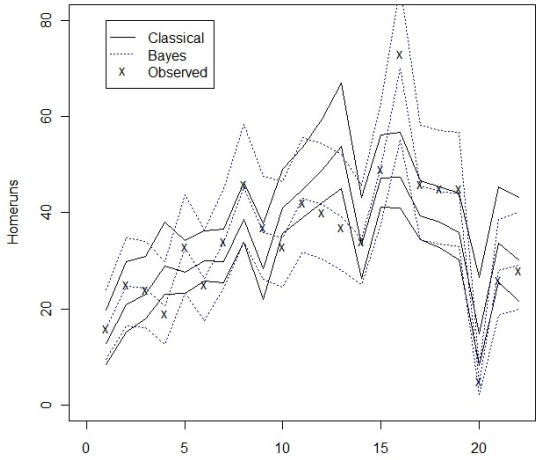

Figure 4.9: Baseball: Comparison of classical method, using at-bats, and and model (4.1) for Bonds.

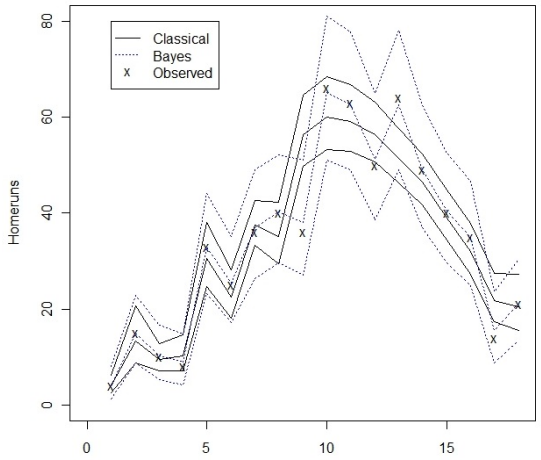

Figure 4.10: Baseball: Comparison of classical method, using at-bats, and and model (4.1) for Sosa.

CHAPTER FIVE

Simulation Studies

We are often interested in how well a proposed model tracks the truth in any given situation. Unfortunately, the true relationship in any real application is unknown to the researcher and, in fact, is what the researcher is attempting to model. Under controlled experiments where we know the truth, we are able to ascertain the model's properties and behavior. This can be done under computer simulations.

In this chapter we consider some limited simulations where we study the properties of model (4.1) under a controlled environment where we know the truth. Model (4.1) is designed to track what we have called inflation factors, otherwise known as relative risks. Using the chosen fixed parameters and the covariate values, we assumed a constant risk, $r = 0.05$, across the entire region. For perspective the estimated constant risk of the Waco Habitat burglary data set was about 0.04. We then calculate the constant mean, μ_r, as the product of the total households at risk and the constant risk r. We then adjust μ_r up or down, based on the calculated inflation factor, λ_i, with the relationship $log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ and $\mu_i = \mu_r \lambda_i$. Then for a single run of the simulation we generated one count y_i for each of the 30 regions using a Poisson(μ_i) distribution where we assumed independence between regions for simplicity. We repeated this process for 1000 total runs for each pair of covariates (β_1, β_2) using the seeds 1 through 1000. Generated data was from R and the model was run in WinBUGS using the R2WinBUGS package. The simulations were run on 64 bit Windows Vista system with 8GB of RAM and a 2.4ghz Intel Quad Core processor. The simulations took on average 33 to 34 hours to run when using four sessions of R at the same time and running 250 total simulations each for a sum of 1000.

The generated model output from the simulations were based on model (4.1) which assumed that the data were generated with spatial correlation. This was done as it is often the case with spatial data that the researcher is not certain if there is or is not spatial correlation, although there are tests to help with this decision. We chose to model the independently generated data using spatial correlation to determine how the model behaves as stated under the simplest case of independence. Furthermore, the generated data are generated as a Poisson count where we satisfy the equality of mean and variance. However, model (4.1) assumes that there is over-dispersion in the data. We will discuss the results of this shortly in section 5.3.

We ran a total of 39 simulations, each as a pair (β_1, β_2), using the centered covariate values for both renter and poverty from the Waco Habitat burglary application found in section 3. We also used the spatial grid associated with the Waco Habitat burglary study. We arbitrarily chose values for the β_1 and β_2 parameters within and on the boundaries of the unit square. We set the intercept parameter β_0 at zero and used chosen combinations of the values 1, .5, .25, .15, and 0, including their negatives, except the pairs $(1, 1)$, $(-1, 1)$, $(1, -1)$, and $(-1, -1)$ which were not run. We used all possible combinations of -0.15, 0, and 0.15; all combinations of -0.25, 0, 0.25; all combinations of -1, -0.5, 0, 0.5, and 1 with the exceptions of those listed above. For each chosen parameter pair we present in the following sections properties determined by these simulation studies.

5.1 Coverage

We first consider how often model (4.1) produced credible sets that covered the covariates. In our experiment we found that for each of the 90%, 95%, and 99% credible sets for the β_1 and β_2 parameters had coverage of at least 0.999 or 1. Table 5.1 displays the coverage rates for the β_1 parameter.

Table 5.1: Credible set coverage for β_1.

(β_1, β_2)	-0.15	0	0.15	-0.15	0	0.15
-0.15	1	0.999	1	1	1	0.999
0	1	1	1	1	1	1
0.15	0.999	1	1	1	1	1

The first column in Table 5.1 represents the values for β_1, the top row represents the values for β_2, columns four through six are for the 90% credible sets, and columns seven through nine are for the 99% credible sets. We notice that only one covariate pair $(\beta_1, \beta_2) = (-0.15, 0.15)$ for the 99% credible sets has coverage of 0.999 for the β_1 parameter, whereas the coverage for the β_2 parameter was observed at 1 (not displayed). All the observed 95% credible sets achieved coverage of unity, for both β_1 and β_2, for every studied covariate pair. However, the same cannot be said for the observed 90% credible sets.

A total of three covariate pairs had observed coverage of 0.999. For the β_1 parameter the two observed coverages of 0.999 where for the covariate pairs $(-0.15, 0)$ and $(0.15, -0.15)$ displayed in Table 5.1. For the β_2 parameter the single observed coverage of 0.999 was for the covariate pair $(0.25, 0)$ which is not displayed. All other pairs had achieved coverage of unity.

We now take notice on the β_0 parameter which we fixed at zero for the simulation study. What we found is that the coverage for the covariate effects was nearly unity for all covariate pairs for both β_1 and β_2. We do not observe this for the $\beta_0 = 0$ case. As a matter of fact, we observe significantly changing coverage throughout the covariate pairs. We discuss this now and display the information in the following tables.

We first begin with the 90% credible set. For the covariate pair $(\beta_1, \beta_2) = (-0.5, 0.5)$ we see coverage of 0.05 whereas when $(\beta_1, \beta_2) = (-0.5, -0.5)$ we see coverage of 0.976. We also make mention that as the covariate pairs reach the extremes, the corners in Table 5.2, the coverage approaches zero. This is due to

the considerable negative bias we find in the estimates for β_0 which we will discuss further in section 5.3. Tables 5.3 and 5.4 display similar information but for the 95% and 99% credible sets respectively which shows an increasing coverage. This is expected as the credible set widths increase.

5.2 Power

In this section we discus the power of the model to detect a covariate effect when indeed one exists. That is, we present the percentage of times the obtained credible sets exclude the value zero when indeed the covariate value is not zero. Since we fixed the value for β_0 at zero, we do not display any such results for this parameter.

We first start with the 90% credible set for the parameter β_1 which Table 5.5 displays the observed results from our simulation. Notice that for the covariate pair $(\beta_1, \beta_2) = (-0.25, 0.25)$ we observe a power of 0.995 whereas for the covariate pair $(-0.15, -0.15)$ we observe a power of 0.383. When $\beta_1 = 0$ then we do not display any information, represented by the value NA, since power is defined as the probability of correctly rejecting the null hypothesis in the frequentist context. Tables 5.6 and 5.7 display similar information where the power tends to decrease, relative to Table 5.5, as the credible sets are wider and have more of a chance to include the value of zero.

We now turn our attention to the power for the β_2 parameter. Tables 5.8, 5.9, and 5.10 display this information in the same format as above.

Table 5.2: 90% Credible set coverage for β_0.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0			0			0	
-0.5	0	0.976			1			0.05	0
-0.25			1		1		1		
-0.15				1	1	1			
0	1		1	1	1	0.94	0.034	0	0
0.15				1	0.994	0.006			
0.25			1		0.346		0		
0.5	1	1			0			0	0
1		0			0			0	

Table 5.3: 95% Credible set coverage for β_0.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0			0.003			0	
-0.5	0	0.999			1			0.288	0
-0.25			1		1		1		
-0.15				1	1	1			
0	1		1	1	1	0.997	0.217	0	0
0.15				1	1	0.081			
0.25			1		0.782		0		
0.5	1	1			0			0	0
1		0			0			0	

Table 5.4: 99% Credible set coverage for β_0.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0			0.371			0.104	
-0.5	0	1			1			0.956	0
-0.25			1		1		1		
-0.15				1	1	1			
0	0.999		1	1	1	1	0.94	0	0
0.15				1	1	0.779			
0.25			1		0.999		0		
0.5	0.999	1			0			0	0
1		0			0			0	

Table 5.5: 90% Credible set power for β_1.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		1			1			1	
-0.5	1	1			1			1	1
-0.25			0.982		0.99		0.995		
-0.15				0.383	0.398	0.405			
0	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.15				0.443	0.466	0.453			
0.25			0.997		0.998		1		
0.5	1	1			1			1	1
1		1			1			1	

Table 5.6: 95% Credible set power for β_1.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		1			1			1	
-0.5	1	1			1			1	1
-0.25			0.9		0.917		0.952		
-0.15				0.116	0.126	0.126			
0	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.15				0.155	0.149	0.163			
0.25			0.962		0.984		0.986		
0.5	1	1			1			1	1
1		1			1			1	

Table 5.7: 99% Credible set power for β_1.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		1			1			1	
-0.5	1	1			1			1	1
-0.25			0.351		0.409		0.485		
-0.15				0.002	0.004	0.001			
0	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.15				0.004	0.002	0.002			
0.25			0.531		0.531		0.56		
0.5	1	1			1			1	1
1		1			1			1	

Table 5.8: 90% Credible set power for β_2.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		1			NA			1	
-0.5	1	1			NA			1	1
-0.25			0.979		NA		0.995		
-0.15				0.404	NA	0.433			
0	1		0.986	0.408	NA	0.444	0.989	1	1
0.15				0.413	NA	0.418			
0.25			0.988		NA		0.988		
0.5	1	1			NA			1	1
1		1			NA			1	

Table 5.9: 95% Credible set power for β_2.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		1			NA			1	
-0.5	1	1			NA			1	1
-0.25			0.905		NA		0.944		
-0.15				0.161	NA	0.15			
0	1		0.915	0.173	NA	0.148	0.94	1	1
0.15				0.171	NA	0.156			
0.25			0.917		NA		0.94		
0.5	1	1			NA			1	1
1		1			NA			1	

Table 5.10: 99% Credible set power for β_2.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		1			NA			1	
-0.5	1	1			NA			1	1
-0.25			0.422		NA		0.518		
-0.15				0.004	NA	0.006			
0	1		0.438	0.004	NA	0.005	0.492	1	1
0.15				0.004	NA	0.005			
0.25			0.424		NA		0.402		
0.5	1	1			NA			1	1
1		1			NA			1	

5.3 Mean Estimates

In this section we consider the mean of the estimates for each of the parameters β_0, β_1, and β_2. We also provide the associated standard deviation for the distribution of the observed Bayesian point estimates, based on squared error loss, returned from model (4.1). For each of the 1000 simulated data sets, we obtained a Bayesian point estimate using the posterior mean of the posterior distribution and display here the mean of those estimates along with the sample standard deviation of those means. Table 5.11 displays the mean for the resulting posterior mean estimates for β_0 and Table 5.12 displays the associated standard deviation.

What we notice in Table 5.11 is a significant negative bias when both β_1 and β_2 get positive and large. Furthermore, this bias is also evident in the upper left portion of the table where β_1 and β_2 get negative and large, although the bias does not appear to be as significant. For example, when the covariate pair $(\beta_1, \beta_2) = (.5, .5)$ the mean of the distribution of $\widehat{\beta_0}$ is -0.832 with a estimated standard deviation of 0.023.

This could be a result of no over-dispersion in the generated data and since the model places the log-linear relationship on the mean of the Poisson rates, assuming that the Poisson rates are random, rather than the rates themselves model (4.1) uses the β_0 parameter to correct for this. That is to say, we observe the negative bias in the β_0 term as a result of the fact the data are distributed Poisson, with equality of mean and variance, and model (4.1) assumes that there is over-dispersion in the data. When this over-dispersion is not present the parameter β_0 is adjusted downward reducing the variance of the Gamma. This could also be a result of an induced prior or the fact that the used covariates are probably correlated, the poverty and renter terms from the Waco data set.

We notice in Tables 5.13 and 5.15 that model (4.1) provides quite accurate estimates for the parameters β_1 and β_2 respectively. The associated standard deviations of these distributions can be found in Tables 5.14 and 5.16 respectively.

Table 5.11: Mean of Bayesian point estimates $\widehat{\beta_0}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		-0.269			-0.159			-0.167	
-0.5	-0.235	-0.621			-0.012			-0.115	-0.045
-0.25			0.033		0.015		-0.049		
-0.15				0.037	0.001	-0.029			
0	-0.04		0.047	0.031	-0.007	-0.066	-0.118	-0.3	-0.937
0.15				0.004	-0.052	-0.13			
0.25			0.007		-0.095		-0.266		
0.5	0.035	-0.006			-0.262			-0.832	-1.75
1		-0.351			-0.847			-1.649	

Table 5.12: Standard Deviation of Bayesian point estimates $\widehat{\beta_0}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.022			0.018			0.017	
-0.5	0.02	0.015			0.012			0.016	0.022
-0.25			0.009		0.011		0.013		
-0.15				0.009	0.011	0.013			
0	0.013		0.009	0.01	0.012	0.014	0.016	0.019	0.024
0.15				0.012	0.014	0.016			
0.25			0.011		0.015		0.018		
0.5	0.009	0.012			0.018			0.023	0.026
1		0.02			0.024			0.026	

Table 5.13: Mean of Bayesian point estimates $\widehat{\beta_1}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		-0.975			-0.987			-0.993	
-0.5	-0.485	-0.494			-0.497			-0.498	-0.497
-0.25			-0.248		-0.248		-0.247		
-0.15				-0.148	-0.149	-0.148			
0	0.001		0.001	0.002	0.002	0.001	0.001	0.002	0.002
0.15				0.152	0.151	0.151			
0.25			0.251		0.25		0.252		
0.5	0.498	0.501			0.501			0.499	0.497
1		0.997			0.995			0.99	

Table 5.14: Standard Deviation of Bayesian point estimates $\widehat{\beta}_1$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.058			0.051			0.043	
-0.5	0.055	0.049			0.041			0.036	0.031
-0.25			0.04		0.039		0.035		
-0.15				0.038	0.037	0.035			
0	0.046		0.038	0.037	0.035	0.034	0.001	0.031	0.026
0.15				0.035	0.033	0.032			
0.25			0.036		0.033		0.031		
0.5	0.041	0.035			0.032			0.028	0.023
1		0.033			0.029			0.025	

Table 5.15: Mean of Bayesian point estimates $\widehat{\beta}_2$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		-0.501			-0.001			0.499	
-0.5	-0.997	-0.497			0.001			0.499	0.996
-0.25			-0.249		0		0.25		
-0.15				-0.149	0	0.15			
0	-0.994		-0.249	-0.15	0	0.15	0.249	0.496	0.991
0.15				-0.15	0.001	0.149			
0.25			-0.25		0		0.246		
0.5	-0.994	-0.499			-0.003			0.496	0.978
1		-0.498			-0.001			0.487	

Table 5.16: Standard Deviation of Bayesian point estimates $\widehat{\beta}_2$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.041			0.04			0.039	
-0.5	0.045	0.043			0.041			0.039	0.036
-0.25			0.043		0.041		0.038		
-0.15				0.041	0.04	0.039			
0	0.047		0.042	0.041	0.04	0.036	0.038	0.035	0.031
0.15				0.041	0.04	0.038			
0.25			0.041		0.039				
0.5	0.049	0.042			0.038		0.036	0.033	0.028
1		0.041			0.035			0.029	

5.4 Credible Set Widths

In this section we display the mean credible set widths along with their associated standard deviations for each of the parameters β_0, β_1, and β_2 for each of the 90%, 95%, and 99% credible sets. Recall that we kept the sample size constant at one observed y_i for each of the 30 regions for the Waco Police Beat map. We start with the 90% credible set widths for parameter β_0 found in Table 5.17 and the associated standard deviation found in Table 5.18. For example, the covariate pair $(-0.15, 0.15)$ results in a 90% credible set width of 0.176 with associated standard deviation of 0.003 for the β_0 parameter.

Tables 5.23 and 5.24 provide the mean widths for the 95% credible sets and standard deviations respectively for β_0 while Tables 5.29 and 5.30 provide the same information but for the 99% credible set widths. Similar mean width information for β_1 can be found in Tables 5.19, 5.25, and 5.31 while Tables 5.20, 5.26, and 5.32 provide the associated standard deviations. Similar mean width information for β_2 can be found in Tables 5.21, 5.27, and 5.33 while Tables 5.22, 5.28, and 5.34 provide the associated standard deviations.

What we observe in these tables is that the credible set widths and associated standard deviations are pretty stable until the parameters move to the extremes of the table. Even in that scenario there appears to be only a moderate increase in width and associated standard deviation.

5.5 Discussion

In our simulation studies we have found that model (4.1) provides coverage of near unity, reasonable power for detecting a covariate effect when one is indeed present while at the same time providing quite accurate point estimates with the exception for the β_0 parameter which shows considerable negative bias. It is our hypothesis that this negative bias is a result of the fact that our model assumes the

Table 5.17: Mean of 90% credible set widths for $\widehat{\beta_0}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.213			0.197			0.189	
-0.5	0.205	0.186			0.177			0.181	0.201
-0.25			0.174		0.174		0.177		
-0.15				0.173	0.174	0.176			
0	0.184		0.172	0.172	0.175	0.178	0.181	0.194	0.23
0.15				0.174	0.177	0.182			
0.25			0.174		0.18		0.19		
0.5	0.176	0.175			0.189			0.225	0.288
1		0.195			0.227			0.283	

Table 5.18: Standard deviation of 90% credible sets widths for $\widehat{\beta_0}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.006			0.005			0.004	
-0.5	0.005	0.004			0.003			0.003	0.004
-0.25			0.003		0.003		0.003		
-0.15				0.003	0.003	0.003			
0	0.003		0.003	0.003	0.003	0.003	0.003	0.005	0.004
0.15				0.003	0.003	0.003			
0.25			0.003		0.003		0.003		
0.5	0.003	0.003			0.003			0.005	0.005
1		0.004			0.005			0.006	

Table 5.19: Mean of 90% credible set widths for $\widehat{\beta_1}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.426			0.386			0.356	
-0.5	0.408	0.367			0.336			0.32	0.319
-0.25			0.331		0.321		0.314		
-0.15				0.321	0.316	0.312			
0	0.359		0.318	0.315	0.311	0.308	0.307	0.308	0.33
0.15				0.311	0.308	0.307			
0.25			0.312		0.308		0.308		
0.5	0.333	0.316			0.311			0.325	0.375
1		0.329			0.339			0.375	

Table 5.20: Standard deviation of 90% credible sets widths for $\widehat{\beta}_1$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.012			0.011			0.009	
-0.5	0.01	0.009			0.007			0.006	0.006
-0.25			0.007		0.006		0.006		
-0.15				0.006	0.006	0.006			
0	0.007		0.006	0.006	0.006	0.006	0.006	0.008	0.005
0.15				0.006	0.006	0.006			
0.25			0.006		0.006		0.005		
0.5	0.006	0.006			0.006			0.009	0.005
1		0.007			0.01			0.011	

Table 5.21: Mean of 90% credible set widths for $\widehat{\beta}_2$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.34			0.327			0.32	
-0.5	0.36	0.337			0.321			0.313	0.322
-0.25			0.325		0.317		0.313		
-0.15				0.32	0.316	0.314			
0	0.357		0.322	0.319	0.315	0.314	0.313	0.316	0.345
0.15				0.318	0.316	0.315			
0.25			0.321		0.317		0.318		
0.5	0.354	0.33			0.324			0.342	0.398
1		0.345			0.355			0.395	

Table 5.22: Standard deviation of 90% credible sets widths for $\widehat{\beta}_2$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.006			0.008			0.008	
-0.5	0.007	0.006			0.006			0.006	0.006
-0.25			0.006		0.006		0.006		
-0.15				0.006	0.006	0.006			
0	0.007		0.006	0.006	0.006	0.006	0.006	0.008	0.006
0.15				0.006	0.006	0.006			
0.25			0.006		0.006		0.006		
0.5	0.007	0.006			0.006			0.009	0.006
1		0.007			0.01			0.012	

Table 5.23: Mean of 95% credible set widths for $\widehat{\beta_0}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.255			0.236			0.227	
-0.5	0.246	0.223			0.211			0.217	0.241
-0.25			0.208		0.208		0.212		
-0.15				0.207	0.208	0.211			
0	0.22		0.206	0.207	0.209	0.214	0.217	0.232	0.275
0.15				0.209	0.212	0.218			
0.25			0.208		0.215		0.227		
0.5	0.211	0.21			0.226			0.27	0.345
1		0.233			0.272			0.339	

Table 5.24: Standard deviation of 95% credible sets widths for $\widehat{\beta_0}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.007			0.006			0.005	
-0.5	0.006	0.005			0.004			0.004	0.005
-0.25			0.004		0.004		0.004		
-0.15				0.004	0.004	0.004			
0	0.004		0.004	0.004	0.004	0.004	0.004	0.006	0.005
0.15				0.004	0.004	0.004			
0.25			0.004		0.004		0.004		
0.5	0.004	0.004			0.004			0.006	0.006
1		0.005			0.007			0.008	

Table 5.25: Mean of 95% credible set widths for $\widehat{\beta_1}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.509			0.462			0.427	
-0.5	0.489	0.44			0.403			0.384	0.384
-0.25			0.397		0.385		0.377		
-0.15				0.385	0.379	0.375			
0	0.43		0.382	0.378	0.373	0.37	0.369	0.37	0.396
0.15				0.373	0.37	0.369			
0.25			0.374		0.37		0.37		
0.5	0.399	0.379			0.374			0.392	0.452
1		0.395			0.408			0.452	

Table 5.26: Standard deviation of 95% credible sets widths for $\widehat{\beta}_1$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.014			0.013			0.011	
-0.5	0.012	0.01			0.008			0.008	0.007
-0.25			0.008		0.008		0.007		
-0.15				0.007	0.007	0.007			
0	0.009		0.008	0.007	0.007	0.007	0.007	0.01	0.006
0.15				0.007	0.007	0.007			
0.25			0.007		0.007		0.006		
0.5	0.008	0.007			0.007			0.012	0.007
1		0.008			0.013			0.015	

Table 5.27: Mean of 95% credible set widths for $\widehat{\beta}_2$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.408			0.392			0.384	
-0.5	0.431	0.404			0.384			0.376	0.387
-0.25			0.389		0.38		0.375		
-0.15				0.384	0.379	0.376			
0	0.427		0.386	0.382	0.378	0.376	0.376	0.379	0.415
0.15				0.382	0.379	0.378			
0.25			0.385		0.38		0.382		
0.5	0.424	0.396			0.388			0.411	0.48
1		0.414			0.427			0.476	

Table 5.28: Standard deviation of 95% credible sets widths for $\widehat{\beta}_2$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.007			0.01			0.01	
-0.5	0.009	0.008			0.007			0.007	0.007
-0.25			0.007		0.007		0.007		
-0.15				0.007	0.007	0.007			
0	0.009		0.007	0.007	0.007	0.007	0.007	0.01	0.007
0.15				0.007	0.007	0.007			
0.25			0.007		0.007		0.007		
0.5	0.009	0.008			0.007			0.012	0.008
1		0.008			0.013			0.016	

Table 5.29: Mean of 99% credible set widths for $\widehat{\beta_0}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.339			0.314			0.301	
-0.5	0.326	0.296			0.281			0.289	0.321
-0.25			0.277		0.277		0.283		
-0.15				0.275	0.276	0.28			
0	0.293		0.274	0.275	0.279	0.284	0.29	0.309	0.366
0.15				0.278	0.283	0.29			
0.25			0.278		0.286		0.302		
0.5	0.28	0.279			0.301			0.358	0.46
1		0.31			0.362			0.452	

Table 5.30: Standard deviation of 99% credible sets widths for $\widehat{\beta_0}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.01			0.009			0.009	
-0.5	0.008	0.007			0.006			0.006	0.007
-0.25			0.006		0.005		0.006		
-0.15				0.006	0.005	0.006			
0	0.006		0.005	0.005	0.006	0.006	0.006	0.009	0.008
0.15				0.006	0.006	0.006			
0.25			0.006		0.006		0.006		
0.5	0.006	0.006			0.006			0.01	0.009
1		0.007			0.012			0.012	

Table 5.31: Mean of 99% credible set widths for $\widehat{\beta_1}$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.677			0.616			0.569	
-0.5	0.651	0.586			0.538			0.513	0.515
-0.25			0.529		0.514		0.504		
-0.15				0.514	0.507	0.501			
0	0.572		0.51	0.505	0.499	0.496	0.494	0.496	0.534
0.15				0.499	0.496	0.494			
0.25			0.5		0.495		0.497		
0.5	0.533	0.507			0.502			0.527	0.612
1		0.53			0.549			0.611	

Table 5.32: Standard deviation of 99% credible sets widths for $\widehat{\beta}_1$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.019			0.019			0.017	
-0.5	0.018	0.015			0.012			0.011	0.011
-0.25			0.012		0.011		0.011		
-0.15				0.011	0.011	0.01			
0	0.013		0.011	0.011	0.011	0.011	0.011	0.017	0.01
0.15				0.011	0.011	0.011			
0.25			0.011		0.011		0.01		
0.5	0.012	0.011			0.011			0.021	0.012
1		0.012			0.022			0.028	

Table 5.33: Mean of 99% credible set widths for $\widehat{\beta}_2$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.545			0.523			0.512	
-0.5	0.576	0.538			0.513			0.502	0.519
-0.25			0.519		0.508		0.501		
-0.15				0.512	0.506	0.502			
0	0.569		0.514	0.51	0.505	0.502	0.502	0.507	0.558
0.15				0.51	0.506	0.506			
0.25			0.514		0.508		0.512		
0.5	0.565	0.528			0.52			0.552	0.648
1		0.553			0.574			0.642	

Table 5.34: Standard deviation of 99% credible sets widths for $\widehat{\beta}_2$.

(β_1, β_2)	-1	-0.5	-0.25	-0.15	0	0.15	0.25	0.5	1
-1		0.011			0.016			0.016	
-0.5	0.013	0.011			0.011			0.011	0.011
-0.25			0.011		0.011		0.011		
-0.15				0.01	0.011	0.011			
0	0.013		0.011	0.011	0.011	0.011	0.011	0.017	0.011
0.15				0.011	0.011	0.01			
0.25			0.011		0.011		0.011		
0.5	0.013	0.012			0.011			0.02	0.013
1		0.012			0.022			0.029	

the variance of the observations is greater than that of the mean, however, under the simulations we assumed both independence and equality of mean and variance for the generated data. Therefore, it appears that model (4.1) adjusts downward the mean of the assumed Gamma distributed Poisson rates through a negative bias on the β_0 parameter to account for this lack of over-dispersion. The credible set widths, and associated standard deviations, seem rather stable for all the covariate pairs chosen until the values of β_1 and β_2 move to extreme values.

In short, the model behaved quite well under our limited simulation studies. Further simulations would be recommended to study the case where spatial correlation and over-dispersion are also present in the observed y_i.

CHAPTER SIX

Discussion

Public health researchers, criminologists, biologists, ecologists, and many other forms of science rely on observed count data as opposed to controlled experiments to construct hypotheses about a system under study. It is desirable to determine a cause and effect relationship from data, unfortunately in the absence of a controlled experiment we are unable to do this effectively and efficiently most of the time.

In this dissertation we have studied the Poisson regression model outlined in McBride (2006), changed the parameterization and proposed prior distributions which improved model behavior when using the software WinBUGGS, and reanalyzed the original baseball and habitat burglary data using the newly proposed model. We then studied the proposed model (4.1) under the simplified assumptions of independence between regions, no over-dispersion, and no spatial correlation. As unrealistic as that scenario is in practice we found that model (4.1) provided stable credible set widths and quite accurate point estimates for the parameter effects while the intercept term proved to be heavily biased toward large negative values.

There are many potential explanations for the observed bias in the β_0 term. On explanation is that the generated data were not over-dispersed. Recall that the log-linear relationship is placed on the mean of the assumed gamma distributed inflation factors. Since the variance of the gamma distribution is a linear multiple of its mean, $V(X) = \frac{\alpha}{\beta^2} = \frac{E(X)}{\beta}$, model (4.1) seems to down play the variability in this distribution by correcting the mean downward through the β_0 term. This effectively reduces the variability in the gamma distribution, which in the absence of over-dispersion we would not need this level of the hierarchy at all. That is to say, if we knew *a-priori* that the data were indeed Poisson and satisfied the equality of

mean and variance we would expect a near exact log-linear relationship as opposed to a log-linear mean relationship. It is also possible that this bias is due to the fact that we did not include an interaction term and used covariates that, in reality, are correlated; however, the generated data did not include an interaction term. It is also possible that the induced prior on the actual mean of the Gamma distribution is causing the bias.

The observed autocorrelation problem in the chains was discussed with the assumption that we used independent priors for the covariate terms. As suggested by some researchers including Ntzoufras (2009) we could have adopted a multivariate normal prior for these terms to have helped alleviate this issue. This would require prior information on the correlation structure of the covariate terms. In our habitat burglary example we used both poverty and the number of renters as covariates and there is good reason to believe that there would be a positive and moderate to high correlation structure between these two covariates. However, in the absence of such *a priori* information it using independent priors may be desirable since the *a priori* assumption of independence does not force independence in the posterior distributions of the covariate terms.

We did consider this option, using (1.15), for the habitat burglary data at the end of our research on this issue and it did help considerably with the autocorrelation in the chains. We chose to use a multivariate normal prior for just the renter and poverty terms and assumed independence between the area and intercept terms; the correlation value we chose was 0.75 and we did not spend much time considering this model structure instead focusing on the independence structure, nor did we present this approach in this work.

It is possible that the observed decrease in thinning rate using our suggested structure was due to a number of different reasons. These may include an interaction between the specific data available and the computer platform were were using, the

specific chosen beta distribution, the limited support of the beta distribution, or several other factors. We seem to have stumbled on an example where we have better performance using our proposed model as opposed to 1.15. Therefore, we believe that the current parameterization seemed to have behaved better in this application and may show promise in out performing 1.15 in general but we have not tested this assertion. For reference the machines used in this work were three Windows Vista operating systems with 8GB of RAM with 2.4 Intel Quad Core processors.

Another observation we have made concerning our proposed model was that the computer run time was decreased and that the time taken per update using WinBUGS was overall much faster than with (1.15). Although we did not test this assertion this was observed for the Waco data set as well as a current work using the entire State of Texas county map under similar assumptions. The Waco data set was run with only a few seconds difference while using the entire State of Texas county map was significantly faster with (4.1).

As shown throughout the body of this work, model (4.1) has shown adaptability to many different applications including spatial disease mapping. This allows for great flexibility in the possible areas that we envision the model to be used. Furthermore, although we have only presented two applications, we believe the model has performed well both in application and under simulations. Although further simulation studies are warranted and greatly recommended we believe that our new model (4.1) shows real promise to applications in public health data and any situation where over/under-dispersion may be suspected.

BIBLIOGRAPHY

Academies, T. N. (2005), *Firearms and Violence: A Critical Review*, National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; 202-334-3313: The National Academies Press, Washington DC.

Albert, J. (1992), "A Bayesian Analysis of a Poisson Random Effects Model for Homerun Hitters," *blkhas*, 12, 123–124.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC.

Berkson, J. O. (1950), "Are there two regressions?" *Journal of the American Statistician*, 45, 164–180.

Bortkiewicz, L. v. (1898), *Das Gesetz de Kleinen Zahlen*, Leipzig, Teubner.

Bramby, T., Orme, C., and Treble, J. (1991), "Worker Absenteeism: An Analysis Using Micro Data," *Economic Journal*, 101, 214–229.

Bratcher, T. L. and Bhalla, P. (1974), "On the Properties of an Optimal Selection Procedure," *Communications in Statistics*, 3, 191–196.

Cameron, A. and Trivedi, P. (1986), "Econometric Models Based on Count Data: Comparisons and Applications of some Estimators," *Journal of Applied Econometrics*, 1, 29–53.

Cameron, C. A. and Trivedi, P. K. (1998), *Regression Analysis of Count Data (7th Printing 2008)*, Cambridge University Press.

Carroll, Raymond J.; Ruppert, D. S. L. A. C. C. M. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective Second Edition*, Chapman and Hall/CRC Press.

Cochran, W. G. (1968), "Errors of Measurement in Statistics," *Technometrics*, 10, 637–666.

Cressie, N. A. C. (1993), *Statistics for Spatial Data: Revised Edition*, John Wiley & Sons, Inc.

Dellaportas, P. and Stephens, D. A. (1995), "Bayesian Analysis of Errors-in-Variables Regression Models," *Biometrics*, 51(3), 1085–1095.

Eicker, F. (1967), *Limit Theorems for Regressions with Unequal and Dependent Errors*, University of California Press, pp. 59–82.

F., B. H. (1948), "Sunlight as a Causal Factor in Cancer of the Skin of Man," *Journal of the National Cancer Institute*, 9, 247–258.

Feinberg, S. (1981), "Deciding What and Whom to Count," in *The NCS Working Papers: volume 1.*, ed. Lehnan, R.G.; Skogan, W., U.S. Department of Justice, vol. 1.

Feinstein, J. (1989), "The Safety Regulation of U.S. Neuclear Power Plants: Violations, Inspections, and Abnormal Occurrences," *Journal of political Economy*, 97, 115–154.

— (1990), "Detection Controlled Estimation," *Journal of Law and Economics*, 33, 233–276.

Fuller, W. A. (1987), *Measurement Error Models*, Wiley.

Galton, F. (1886), "Regression Towards Mediocrity in Hereditary Stature," *The Journal of the Anthropological Institute of Great Britian and Ireland*, 15, 246–263.

Gamerman, D. and Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference (second edition)*, Chapman & Hall/CRC.

Ghosh, M. and Gelfand, A. (2000), *"Generalized Linear Models: A Bayesian View" in Generalized Linear Models: A Bayesian Perspective*, Marcel Dekker.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo In Practice*, Chapman & Hall/CRC.

Gourieroux, C., Monfort, A., and Trognon, A. (1984a), "Pseudo Maximum Likelihood Methods: Applications to poisson Models," *Econometrica*, 52, 701–720.

— (1984b), "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700.

Gustafson, P. (2004), *Measurement Error and Misclassification in Statistics and Epidemiology*, Chapman and Hall/CRC Press.

Hamilton, C., Bratcher, T. L., and Stamey, J. D. (2008), "Bayesian subset selection approach to ranking normal means," *Journal of Applied Statistics*, 35 No. 8, 847–851.

Johansson, P. and Palme, M. (1996), "Do Economics Incentives Affect Work Absence: Empirical Ecidence Using Swedish Micro Data," *Journal of Public Economics*, 59, 195–218.

Johnson, N., Kotz, S., and Kemp, A. W. (1992), *Univariate Distributions: second edition*, John Wiley: New York.

Jordan, P., Brubacher, D., Tsugane, S., Tsubono, Y., Gey, K., and Moser, U. (1997), "Modeling Mortality Data from a Multi-Center Study in Japan by Means of Poisson Regression with Errors in Variables," *International Journal of Epidemiology*, 26, 501–507.

Kutran, K. (1975), "Bayesian Corrections for Visibility Bias in Population Estimates," Ph.D. thesis, University of Southwestern Louisiana.

Lawson, A. B. (2009), *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, CRC Press.

Liska, A. E., Sanchirico, A., and Reed, M. D. (1988), "Fear of Crime and Constrained Behavior: Specifying and Estimating a Reciprocal Effects Model," *Social Forces*, 66, 827–837.

McBride, J. J. (2006), "Conjugate Hierarchical Models for Spatial Data: An Application of an Optimal Selection Procedure," Ph.D. thesis, Baylor University.

McGlothlin, A., Stamey, J., and Seaman, J. W. J. (2008), "Binary Regression with Misclassified Response and Covariate Subject to Measurement Error: a Bayesian Approach," *Biometrical Journal*, 50, 123–134.

McInturff, P., Johnson, W., Cowling, D., and Gardner, I. (2004), "Modeling risk when binary outcomes are subject to error." *Statistics in Medicine*, 23 no. 7, 1095–1109.

Ntzoufras, I. (2009), *Bayesian Modeling Unsing WinBUGS*, John Wiley & Sons, Inc.

Paulino, C., Soares, P., and Neuhaus, J. (2003), "Binomial regression with misclassification." *Biometrics*, 59 no. 3, 670–675.

Plam, T. (1890), "The geographical distribution and aetiology of rickets." *Practitioner*, 45, 270–279,321–342.

Poisson, S. D. (1837), *Recherches sur la probabilité des Jugements en Matière Criminelle et en Matière Civile*, Paris, Bachelier.

Pridemore, W. A. (2008), "A methodological Addition to the Cross-National Empirical Literature on Social Structure and Homicide: A First Test of the Poverty Homicide Thesis," *Criminology*, 46 No. 1, 133–154.

Reeves, G. K., Cox, D. R., Darby, C., and Whitley, E. (1998), "Some Aspects of Measurement Error in Explanatory Variables for Continuous and Binary Regression Models." *Statistics in Medicine*, 17, 2157–2177.

Richardson, S. and Gilks, W. S. (1993), "Conditional Independence Models for Epidemiological Studies with Covariate Measurement Error," *Statistics in Medicine*, 24(2), 1703–1722.

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods (second edition)*, Springer.

Robinson, P. (1987), "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.

Roy, S., Banerjee, T., and Maiti, T. (2005), "Measurement Error Model for Misclassified Binary Responses," *Statistics in Medicine*, 24(2), 269–283.

Russer, J. (1991), "Workers' Compensation and Occupational Injuries and Illnesses," *Journal of Labor Economics*, 9, 325–350.

Schabenberger, O. and Gotway, C. A. (2005), *Statistical Methods for Spatial Data Analysis*, Chapman and Hall/CRC.

Schneider, A. (1981), "Differences Between Survey and Police Information about Crime," in *The NCS Working Papers: volume 1.*, U.S. Department of Justice.

Schwartz, J. and Coull, B. A. (2003), "Control for Confounding in the Presense of Measurement Error in Hierarchical Models," *Biostatistics*, 4 No. 4, 539–553.

Snow, J. (1854), *On the Mode of Communication of Cholera (2nd ed.)*, London: Churchill Livingstone.

Solow, A. (1993), "Estimating Record Inclusion Probability," *American Statistician*, 47, 206–209.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit (with discussion)," *Journal of the Royal Statistical Society*, 64, 583–639.

Sposto, R., Preston, D. L., Shimizu, Y., and Mabuchi, K. (1992), "The effect of diagnostic misclassification on non-cancer and cancer mortality dose-response in A-bomb survivors." *Biometrics*, 48:2, 605–617.

Stamey, J. (2000), "A Bayesian Analysis of Poisson Data with Misclassification," Ph.D. thesis, Baylor University.

Stamey, J. D., Bratcher, T. L., and Young, D. M. (2004), "Parameter subset selection and multiple comparisons of Poisson rate parameters with misclassification," *Computational Statistics & Data Analysis*, 45, 467–479.

Stamey, J. D., Young, D. M., and Seaman, J. J. (2007), "A Bayesian approach to adjust for diagnostic misclassification between two mortality causes in Poisson regression." *Statistics in Medicine*, 27, 2440–2452.

Watermann, J., Jankowski, R., and Madan, I. (1994), "Under-reporting of Needlestick Injuries by Medical Students," *Journal of Hospital Infection*, 26, 149–151.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 46, 817–838.

Whittemore, A. S. and Gong, G. (1991), "Poisson Regression with Misclassified Counts: Application to Cervival Cancer Mortality Rates," *Applied Statistics*, 40 No. 1, 81–93.

Whittemore, A. S. and Keller, J. B. (1988), "Approximations for Regression with Covariate Measurement Error." *Journal of the American Statistical Association*, 83(404), 1057–1066.

Yannaros, N. (1993), "Analyzing Incomplete Count Data," *Statistician*, 42, 181–187.