2009-07-07

Zero-Inflated Censored Regression Models: An Application with Episode of Care Data

Jonathan P. Prasad
Brigham Young University - Provo

ZERO-INFLATED CENSORED REGRESSION MODELS: AN APPLICATION WITH EPISODE OF CARE DATA

by

Jonathan P. Prasad

A project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

August 2009

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a project submitted by

Jonathan P. Prasad

This project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	G. Bruce Schaalje, Chair
Date	Dennis L. Eggett
Date	Gilbert W. Fellingham

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the project of Jonathan P. Prasad in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

G. Bruce Schaalje
Chair, Graduate Committee

Accepted for the Department

Scott D. Grimshaw
Graduate Coordinator

Accepted for the College

Thomas W. Sederberg
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

ZERO-INFLATED CENSORED REGRESSION MODELS: AN APPLICATION WITH EPISODE OF CARE DATA

Jonathan P. Prasad

Department of Statistics

Master of Science

The objective of this project is to fit a sequence of increasingly complex zero-inflated censored regression models to a known data set. It is quite common to find censored count data in statistical analyses of health-related data. Modeling such data while ignoring the censoring, zero-inflation, and overdispersion often results in biased parameter estimates. This project develops various regression models that can be used to predict a count response variable that is affected by various predictor variables. The regression parameters are estimated with Bayesian analysis using a Markov chain Monte Carlo (MCMC) algorithm. The tests for model adequacy are discussed and the models are applied to an observed data set.

ACKNOWLEDGEMENTS

I wish to thank the following individuals for making it possible for me to successfully complete this project:

- Professor G. Bruce Schaalje of Brigham Young University, Utah for being my mentor and advisor, and for working with me so diligently and patiently.
- Professor Dennis L. Eggett and Professor D. Rusell Crane, Brigham Young University, Utah, for providing the episode of care data set and the initial analyses of the data.
- Professor Gilbert W. Fellingham, Brigham Young University, Utah, for providing ideas and suggestions that helped cover the fine details of the project.
- My beautiful wife, Nancy, for her support and love, and for keeping our little daughter from getting into the project papers.

CONTENTS

TABLES

Table

FIGURES

Figure

1. INTRODUCTION

Psychosocial therapists have studied the "medical offset effect" of family therapy and other forms of therapy service (Crane and Christenson 2007). The medical offset is defined as the reduction in health care use due to a specific kind of psychosocial intervention. A medical offset is considered to have occurred if the adjusted mean utilization of health care is significantly lower when the psychosocial intervention is present.

Morgan and Scovial (2007) examined medical offset in a large Episode of Care (EoC) data set provided by CIGNA Behavioral Health. The EoC data provided information on health care utilization in the form of the number of visits made by patients to their psychotherapy care provider over a period of time. The EoC data has a count response variable called *number of sessions/visits* and various predictor variables including patient age, therapist type, procedure type, and region. The procedure type indicated the nature of the psychosocial intervention: whether it was individual therapy, family therapy, or a mixture of the two. It was anticipated that involvement of family members would have a negative effect on the number of sessions, and thus would be an example of medical offset due to family therapy.

As part of their study, Morgan and Scovial (2007) looked at the first EoC of a patient during the study period for which they used ordinary Gaussian theory regression and censored Weibull regression to relate the log of the number of sessions to the predictor variables. Both statistical analyses might be suitable for the data set given the fact that the data set has many observations (over 100,000), and given the robust nature of Gaussian regression.

However, one concern with the ordinary Gaussian and censored Weibull models is that they fail to thoroughly address the possible problems of *one-inflation* and *overdispersion* that appear to be present in the count data (Figure 1.1). The Gaussian

regression procedure also overlooks the censored aspect of the EoC data; that is, 18% of the observations are known only as a lower limit. For example, even though the data shows that a patient had five sessions, it is possible in certain known cases that the patient's therapy may have begun before the start of the study or may have continued after the end of the study.

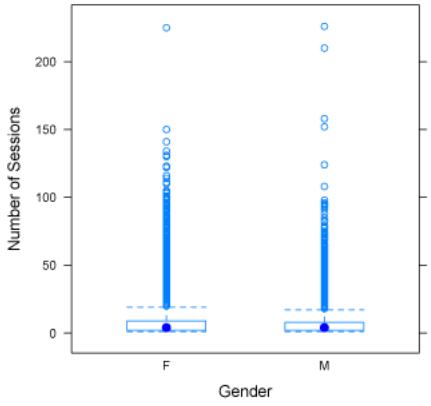

Figure 1.1: Box-plots of numbers of sessions by gender for the CIGNA EoC data. Both groups have long tails and have medians close to one.

We will also look at the first EoC of a patient during the study period, but will adjust the response variable to be Y = number of sessions - 1, so that standard models for count data can be used. The adjusted response, Y, will change the EoC data from being one-inflated to zero-inflated, for which statistical approaches are available.

We propose to analyze the EoC data using a zero-inflated censored Poisson regression model (ZICP), a zero-inflated censored generalized Poisson regression model (ZICGP), and a zero-inflated censored negative binomial regression model (ZICNB) fitted via Bayesian methods to address the aforementioned problems. It is hoped that doing so will produce a better fit and a more realistic model for session counts. This in turn would provide better predictions of utilization for health care insurers. We will also have unbiased evaluations of factors associated with medical offset, not the biased evaluations that result from ignoring censoring.

2. LITERATURE REVIEW

2.1 Models for Count Data

In experimental and observational studies in many disciplines, including public health, economy, psychology, biology, and medicine, regression analysis is used to model dependent variables that describe count data. A frequently used statistical model for count data is the standard Poisson model. If Y follows a Poisson distribution with parameter μ, where $\mu \geq 0$, then

$$f(y) = Pr\{Y = y\} = \frac{e^{-\mu}\mu^{y}}{y!}.$$

The mean, $E(Y)$, and variance, $Var(Y)$, associated with the Poisson distribution are

$$E(Y) = Var(Y) = \mu.$$

A common Poisson regression model for a count response Y_i is

$$Y_i \sim Poisson(\mu_i),$$

where $\mu_i = \exp(\mathbf{x}_i^{'}\boldsymbol{\beta})$, $\boldsymbol{x_i^{'}}$ is a vector of co-variates, and $\boldsymbol{\beta}$ is a vector of parameters. For example, Aitkin, Anderson, Francis, and Hinde (1990) and Renshaw (1994) each fitted the Poisson model to two different sets of auto insurance claims data.

In many real-life applications the counts include more zeros than expected under the Poisson model (Lambert 1992 and Rodrigues-Motta, Gianola, Heringstad, Rosa, Chang 2006). This condition, called *zero-inflation*, occurs because a single Poisson parameter μ is often insufficient to describe the population (Bohning 1994). Essentially, zero-inflation arises when one mechanism generates only zeros and another process generates both zero and nonzero counts (Greene 1994). In general, models that accommodate an excess of zeros are characterized by a parametric structure that models the zero and nonzero mechanisms separately. Zorn (1996) justifies

this approach in terms of a dual regime data generating process: in the first stage, a presence/absence model determines whether the count is zero or possibly nonzero; in the second stage, a count model governs the actual magnitude of the count. Hence the probability density function (pdf) of the model is

$$f(y) \quad = \quad \eta\gamma_0(y) + (1-\eta)f_2(y), \tag{2.1}$$

and the cumulative density function (cdf) is

$$F(y) \quad = \quad \eta + (1-\eta)F_2(y), \tag{2.2}$$

where η is the probability of the zero count in the presence/absence model, $\gamma_0(y)$ is a zero indicator function for the presence/absence model, and $f_2(y)$ is the pdf of a count random variable. This model (Equation 2.1) is often called a mixture model and η is called the mixture proportion. A common mixture model is the zero-inflated Poisson (ZIP) model where $f_2(y)$ is the Poisson pdf (Rodrigues-Motta et al. 2006).

Another common problem in regression with count data is that the data may exhibit greater variation than the Poisson distribution. This is referred to as *overdispersion*. Various models have been proposed in the literature to deal with overdispersion. One such alternative to the standard Poisson model is the negative binomial (NB) model derived from a continuous mixture of Poisson distributions where the mixing distribution of the Poisson counts is a gamma distribution. The NB, also known as a gamma-Poisson (mixture) model, has the pdf:

$$f(y) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}p^k(1-p)^y \tag{2.3}$$

where $k > 0$, $0 \leq p \leq 1$,

$$\begin{aligned} E(Y) &= \frac{k(1-p)}{p}, \text{and} \\ Var(Y) &= \frac{k(1-p)}{p^2} = \frac{E(y)}{p}. \end{aligned}$$

For use in regression, Ismail and Jemain (2007) suggest the reparametrization of Equation 2.3 with

$$\begin{aligned} \phi &= \frac{1}{k}, \text{and} \\ \mu &= \frac{1-p}{\phi p}, \end{aligned}$$

such that

$$f(y) = \frac{\Gamma(y + \frac{1}{\phi})}{\Gamma(y+1)\Gamma(\frac{1}{\phi})} \left(\frac{1}{1+\phi\mu}\right)^{\frac{1}{\phi}} \left(\frac{\phi\mu}{1+\phi\mu}\right)^{y}, \tag{2.4}$$

where ϕ is the overdispersion parameter. The mean of the model is $E(Y) = \mu$ and the variance is $Var(Y) = \mu + \mu^2\phi$ (Ismail and Jemain 2007). As ϕ approaches zero, $f(y)$ approaches the Poisson pdf in the limit. If $\phi > 0$, the variance will exceed the mean, and thus the distribution allows for overdispersion.

To accommodate zero-inflation as well as overdispersion, $f_2(y)$ in Equation 2.1 could be the NB pdf (Equation 2.4). The resulting model is called the zero-inflated negative binomial (ZINB) (Rideout, Hinde, Demetrio 2001).

The generalized Poisson (GP) regression model is another alternative to the standard Poisson model to deal with overdispersion. One parametrization (Famoye and Wang 2004) of the GP is

$$f(y) = \left(\frac{\mu}{1+\alpha\mu}\right)^{y} \frac{(1+\alpha y)^{y-1}}{y!} \exp\left[\frac{-\mu(1+\alpha y)}{1+\alpha\mu}\right]. \tag{2.5}$$

The mean of the GP model is $E(Y_i) = \mu_i$, the variance is $Var(Y_i) = \mu_i(1 + \alpha\mu_i)^2$, and α is the dispersion parameter (Famoye and Wang 2004). When $\alpha = 0$ the model reduces to a standard Poisson. For $\alpha > 0$, the GP model represents count data with overdispersion.

To accommodate zero-inflation as well as overdispersion, $f_2(y)$ in Equation 2.1 could be the GP pdf (Equation 2.5). The resulting model is called the zero-inflated generalized Poisson (ZIGP).

2.2 Censoring

Count data are often censored. Censoring occurs when the starting and ending times of the study result in partial information for some of the counts. For example, in the EoC data some of the patients had either begun their visits prior to the start of study or were continuing past the end of study. Hence, for these patients we only have the information captured during the study period. If no censoring has occurred for the ith patient, the count is $Y_i = y_i$. However, if censoring occurs for the ith patient, we know only that $Y_i \geq y_i$.

To estimate parameters of a count distribution when there is censoring, consider the count distribution $f(y_i; \boldsymbol{\theta})$ and the censoring count distribution $g(z_i; \boldsymbol{\Gamma})$. For every individual i, actual count and censoring count are drawn. Censoring occurs if $z_i < y_i$.

The likelihood function, used in either maximum likelihood or Bayesian estimation of the parameters, is

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta})^{1-\delta_i} S(y_i; \boldsymbol{\theta})^{\delta_i}, \tag{2.6}$$

where δ_i is the censoring indicator such that

$$\delta_i = \begin{cases} 1 & \text{if censored,} \\ 0 & \text{otherwise,} \end{cases}$$

and $S(y_i; \boldsymbol{\theta})$ is the survival function. For a discrete count distribution, the survival function is

$$S(y_i; \boldsymbol{\theta}) = F(Y_i \geq y_i) = 1 - F(y_i - 1; \boldsymbol{\theta}), \tag{2.7}$$

where $F(y_i)$ is the cdf of the count model. Models for count data that are censored will be referred to as ZICP, ZICGP, and ZICNB.

2.3 Bayesian Inference

Bayesian inference is based on Bayes Theorem (Gelman, Carlin, Stern, and Rubin 2004) which states that the conditional probability for events A given B is related to the conditional probability of B given A by

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

To use this theorem for parameter estimation, the idea is to substitute the parameter vector $\boldsymbol{\theta}$ and data $\boldsymbol{y}$ for events A and B to get

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{y})} \tag{2.8}$$

where

$$f(\boldsymbol{y}) = \int \dots \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{2.9}$$

In Equation 2.8, $\pi(\boldsymbol{\theta})$ is called the *prior* pdf of $\boldsymbol{\theta}$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is called the *posterior* pdf.

Bayesian methodology has been around since the 18th century, but extensive use was not possible until recently due to the difficulty in computation. For example, $f(\boldsymbol{y})$ (Equation 2.9), often referred to as the *normalizing constant*, is not easily calculated, even for simple situations.

A common approach in dealing with difficult computation in Bayesian analysis is the use of the Metropolis-Hastings (MH) algorithm. MH is a Markov chain Monte Carlo (MCMC) method used to approximate distributions under simulation. Hastings (1970) generalized the Metropolis algorithm, and the resulting algorithm is commonly referred to as the Metropolis-Hastings algorithm.

Key steps in a Bayesian analysis include:

(1) Specify the prior distribution for $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$ to express uncertainty.

(2) Choose a statistical model or likelihood that describes the distribution of data $\mathbf{y}$ given $\boldsymbol{\theta}$, $f(\boldsymbol{y}|\boldsymbol{\theta})$.

(3) Update uncertainty about $\boldsymbol{\theta}$ by combining information from the prior distribution and the likelihood to calculate the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Usually $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ does not have a closed form, so Markov chain Monte Carlo (MCMC) methods are needed to draw samples from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. The posterior draws generated from the MCMC iterations can then be used to calculate estimates and confidence intervals for the regression parameters or functions of them. The draws can also be combined with the likelihood function to generate predictions and prediction intervals for new individuals.

2.4 Goodness of Fit

A measure of goodness of fit for the proposed model specification in Bayesian analysis is the Bayesian χ^2 statistic as defined by Johnson (2004). This goodness of fit statistic is based on the process of partitioning and binning. For regression data the steps in the process include:

(1) Partition $(0, 1)$ into k intervals (not necessarily equal).

(2) Randomly pick $\tilde{\theta}$ from the posterior draws.

(3) For every $y_i, \boldsymbol{x}_i$ in the data set calculate $F_{\tilde{\theta}}(y_i|\boldsymbol{x}_i)$.

(4) Bin each $F_{\tilde{\theta}}(y_i|\boldsymbol{x}_i)$ using the partition from (1).

(5) Tally the number of $F_{\tilde{\theta}}(y_i|\boldsymbol{x}_i)$ in each partition.

(6) Calculate

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j},$$

where

O_j = the observed number of $F_{\tilde{\theta}}$ in jth partition,

E_j = the expected number of $F_{\tilde{\theta}}$ in jth partition

(E_j = the width of the jth times the total number of observations.)

(7) Repeat the above steps n times.

(8) From the repetitions, calculate $\overline{\chi^2}$.

(9) Compute $p - value$ using the χ^2 distribution with $df = k - 1$.

The Bayesian goodness of fit procedure works well and is easy to compute for noncensored data. However, it can not be directly applied to censoring cases since

$F_{\tilde{\theta}}(y_i|\boldsymbol{x}_i)$ has a different meaning for censored data; it is difficult to assign to an appropriate bin since y_i is a lower limit for actual count. To address this issue Yin (2009) and Cao, Moosman, and Johnson (2007) suggest working only with the non-censored observations. In step (3) they calculate for each noncensored observation

$$\frac{1}{a}\int_0^{y_i}(1-G(z))f(z)dz,$$

where

$$a=\int_0^{\infty}(1-G(z))f(z)dz$$

and $G(z_i)$ is the cdf of the censoring distribution. For discrete random variables, the corresponding quantity would be

$$\frac{1}{a}\sum_{z=0}^{y_i}(1-G(z-1))f(z)dz,$$

where

$$a=\sum_{z=0}^{\infty}(1-G(z-1))f(z)dz.$$

2.5 Model Selection

A widely used statistic for comparing and selecting models in a Bayesian framework is the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, and vander Linde 2002). DIC is a Bayesian alternative to the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Calculation of the DIC in connection with MCMC is trivial as it does not require maximization over the parameter space, like the AIC and BIC.

The deviance, $D(\boldsymbol{\theta})$, is evaluated at the parameter mean $\overline{\boldsymbol{\theta}}$ as

$$D(\overline{\boldsymbol{\theta}}) = -2\log[f(\mathbf{y}|\overline{\boldsymbol{\theta}})].$$

The goodness of fit of the model is summarized as the posterior expectation of the deviance

$$\overline{D}(\theta) = E_{\boldsymbol{\theta}|\boldsymbol{y}}[D].$$

The complexity of the model (the effective number of parameters in the model), p_D, is given by the expected deviance minus the deviance evaluated at the posterior means

$$\begin{aligned} p_D &= E_{\boldsymbol{\theta}|\boldsymbol{y}}[D] - D(E_{\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{\theta}]) \\ &= \overline{D} - D(\overline{\boldsymbol{\theta}}). \end{aligned}$$

These statistics combine to give the overall DIC

$$\begin{aligned} DIC &= p_D + \overline{D} \\ &= 2\overline{D} - D(\overline{\boldsymbol{\theta}}). \end{aligned}$$

Essentially, DIC is a penalized measure of how well the model fits the data. It is a compromise between fit and complexity. A smaller DIC indicates better fit to the data set.

The model with the best DIC and an acceptable Bayesian χ^2 will be the one considered for prediction purposes.

3. DATA DESCRIPTION

The data were obtained from one of the leading health care insurers in the United States, CIGNA Behavioral Health (CIGNA). The use of administrative data for the purposes of compiling aggregate statistics, monitoring trends, and providing information for planning purposes is allowed by the Health Insurance Portability and Accountability Act of 1996 (HIPAA), regulations for protecting personal health care information. The raw claims data received from CIGNA contained over five million claims for the study period, starting January 1, 2001 and ending December 31, 2003. Names and all personal identifying information were removed and a unique and unidentifiable client identification number was added for each patient prior to the data being delivered. In no case was it possible to identify any unique subscriber or provider information from the data provided.

For our analysis we used a sample of the raw data containing only patients who were receiving anxiety-related therapy. An episode of care (EoC) is defined by CIGNA as a series of continuing services for the same patient; it begins with the first service and ends when the individual has had no claims for 90 days or more. Since the data set included the date of each service, it was possible to compute the number of EoCs as well as the number of sessions in that EoC for each patient/client. For this project we used the number of sessions for each patient in the *first* EoC during the study period.

The response variable, Y, is the total number of sessions (for a patient) minus one. Note that, by definition, an EoC must have at least one session. Hence Y has a minimum of zero. The objective of the study is to relate Y to various predictor variables.

We added a censoring indicator variable to the data. If, during the study period, the patient's first visit took place within the first 89 days of the beginning (Jan. 1,

2001) of the study period or within the last 89 days of the end (Dec. 31, 2003) of the study period (Figure 3.1), the observation was categorized as being censored.

Data from all states of the U.S.A. were included in the study. For regional comparison purposes, we created a region variable where the states were summed into six distinct regions: *Northeast* (CT, DE, MA, ME, NH, NJ, NY, PA, RI, VT); *Midwest* (IL, IN, IA, KS, MI, MN, MO, ND, NE, OH, OK, SD, WI); *Pacific* (AK, CA, HI, OR, WA); *South* (AL, AR, DC, FL, GA, KY, LA, MD, MS, NC, SC, TN, VA, WV); and *West* (AZ, CO, ID, MT, NM, NV, TX, UT, WY). Services performed outside the United States were not considered for our analysis.

Other predictor variables included client age (in years), client gender, procedure type (individual, family, or mixed), and therapist license (psychologist, psychiatrist, or social worker).

3.1 Descriptive Statistics

The sample data used for our analysis had 123,163 observations. Each row of the data had the first EoC information for each patient. Of the 123,163 observations, 84,162 were for females and 39,001 were for males. Furthermore, 22,936 had a censored number of sessions (Table 3.1), of which 7,266 were for males, and 15,670 were for females.

Table 3.1: EoC Summary

	EoC	Median no. Sess.	Max. no. Sess.	I.Q.R[a]	Single Sess.[b]	Percent[c](%)
total	123,163	4	226	6	22,897	18.6
non-censored	100,227	4	226	7	17,555	17.5
censored	22,936	3	210	5	5,346	23.3
start censored[d]	9,765	2	42	3	3,441	35.2
end censored[e]	13,171	5	210	10	1,905	14.5

[a] inter quartile range
[b] number of single session in the EoC
[c] percent of single session EoCs of total EoC
[d] censoring between Jan. 1, 2001 - Mar. 30, 2001
[e] censoring between Oct. 4, 2003 - Dec. 31, 2003

The youngest patient was three years old and the oldest was 94 years old. There were 2,932 patients 11 years and younger, 16,007 patients between the ages of 12 and 18, 53,345 patients between the ages of 19 and 40, 47,480 patients between the ages of 41 and 60, and 3,411 patients were 61 years or older (Figure 3.2).

Males and females had similar overall trends in terms of session counts. Both distributions were very long tailed (Figure 3.3). A similar trend in the session count distribution was observed for all the other predictor variables. The median for censored counts was less than the median for noncensored counts. Likewise the mean for the censored counts was less than the mean for the noncensored counts. Most of the EoCs (97,954) involved the individual procedure type.

Figure 3.1: Diagram illustrating type of censoring in the EoC Data. Since we know the lower limits from either type, EoC data has right censoring only.

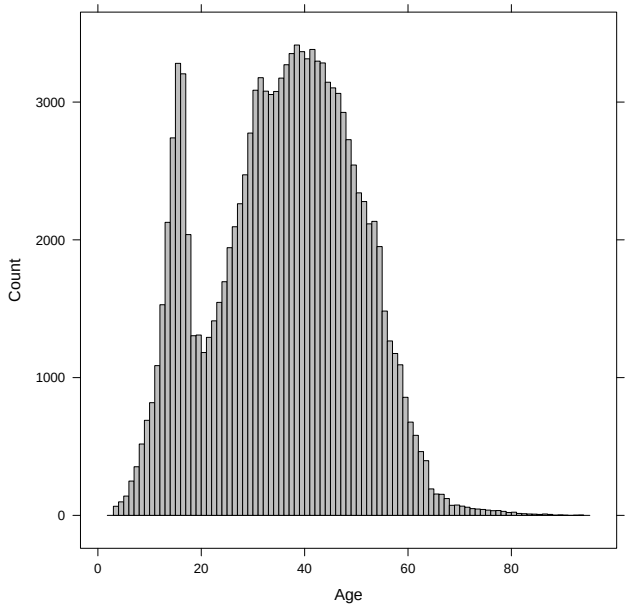

Figure 3.2: Histogram of Ages. Late teens and mid 40s are frequently occurring ages.

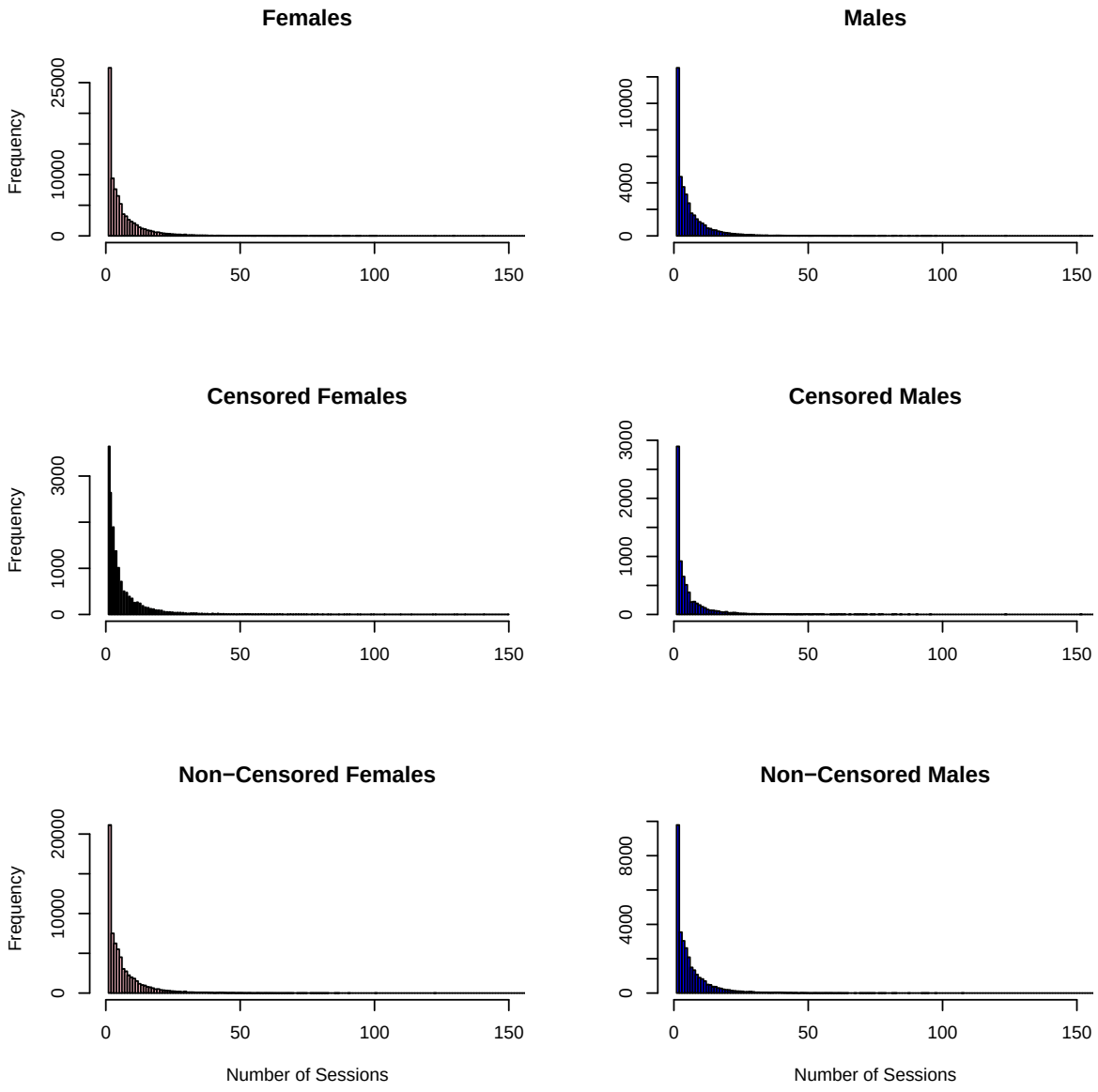

Figure 3.3: Histograms of the Number of Sessions by Gender. The top row of plots are of all the EoCs. The middle row of plots are of censored EoCs, and the bottom row of plots are of noncensored EoCs.

4. ZERO-INFLATED CENSORED MODELS

In this project, we fit a sequence of increasingly complex censored models to the EoC data. The added complexity will help address the issues of zero-inflation, over-dispersion, and censoring in the data. In this section we describe likelihood functions for such models.

4.1 Zero-Inflated Poisson

A zero-inflated model can be expressed as a two-component mixture model where one component has a degenerate distribution at zero and the other is a count distribution (Equation 2.1) such as the Poisson.

Let the response variable, Y_i, be the number of sessions minus one for patient i in the first EoC during the study period, and let n be the number of first EoCs (also the number of patients) in the data. The zero-inflated Poisson (ZIP) distribution is defined as

$$f(y_i) = \eta\gamma_0(y_i) + (1-\eta)\left(\frac{e^{-\mu_i}{\mu_i}^{y_i}}{y_i!}\right) \tag{4.1}$$

where

$$\gamma_0(y_i) = \begin{cases} 1 & \text{if the count belongs to the degenerate distribution} \\ 0 & \text{otherwise} \end{cases}$$

and η is a mixing proportion, $0 \leq \eta \leq 1$. The ZIP regression model is obtained by

defining μ_i as

$$\begin{aligned}
\mu_i &= \exp(\mathbf{x}_i^{'}\boldsymbol{\beta}), \text{where} \\
\mathbf{x}_i^{'} &= \text{a vector of covariates for } EoC_i, \text{ and} \\
\boldsymbol{\beta} &= \text{a } k\text{-dimensional vector of parameters.}
\end{aligned}$$

4.2 Zero-Inflated Censored Poisson

To properly account for the presence of censored EoCs in the data, we will use the survival function (Equation 2.7). The cdf, following Equation 2.2, of an EoC is

$$\begin{aligned}
F(y_i) &= \sum_{s=0}^{y_i} \left[\eta\gamma_0(s) + (1-\eta) \left(\frac{e^{-\mu_i}{\mu_i}^s}{s!} \right) \right] \\
&= \eta + (1-\eta) \sum_{s=0}^{y_i} \left(\frac{e^{-\mu_i}{\mu_i}^s}{s!} \right). \qquad (4.2)
\end{aligned}$$

For a censored EoC we know only that $Y_i \geq y_i$, and since the random variable is discrete, the survival function of a censored EoC is

$$S(y_i) = 1 - F(y_i - 1). \qquad (4.3)$$

Thus, following Equation 2.6, the likelihood function of a zero-inflated censored Poisson (ZICP) regression model is

$$L = \prod_{i=1}^{n} f(y_i)^{1-\delta_i} (S(y_i))^{\delta_i},$$

where

$$\delta_i = \begin{cases} 1 & \text{if censored,} \\ 0 & \text{otherwise.} \end{cases}$$

The log-likelihood function is defined as

$$\ell = (1-\delta_i)\sum_{i=1}^{n}\log\left[f(y_i)\right] + \delta_i\sum_{i=1}^{n}\log\left[S(y_i)\right]. \tag{4.4}$$

4.3 Zero-Inflated Censored Generalized Poisson

For the zero-inflated censored generalized Poisson (ZICGP) regression model we use the same parametrization as Equation 2.5. The log-likelihood of the ZICGP regression model is formulated as Equation 4.4, where

$$f(y_i) = \eta\gamma_0(y_i) + (1-\eta)\left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}\exp\left[\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right] \tag{4.5}$$

and

$$\begin{aligned} F(y_i) &= \sum_{s=0}^{y_i}\left[\eta\gamma_0(s) + (1-\eta)\left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{s}\frac{(1+\alpha s)^{s-1}}{s!}\exp\left[\frac{-\mu_i(1+\alpha s)}{1+\alpha\mu_i}\right]\right] \\ &= \eta + (1-\eta)\sum_{s=0}^{y_i}\left[\left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{s}\frac{(1+\alpha s)^{s-1}}{s!}\exp\left[\frac{-\mu_i(1+\alpha s)}{1+\alpha\mu_i}\right]\right]. \end{aligned} \tag{4.6}$$

When $\alpha = 0$, Equation 4.5 and Equation 4.6 reduce to Equation 4.1 and Equation 4.2, respectively. For $\alpha > 0$, the ZICGP model is suitable for zero-inflated censored count data with overdispersion.

4.4 Zero-Inflated Censored Negative Binomial

For the zero-inflated censored negative binomial (ZICNB) regression model we use the same parametrization as Equation 2.4. The log-likelihood of the ZICNB regression model is formulated as Equation 4.4, where

$$f(y_i) = \eta\gamma_0(y_i) + (1-\eta)\frac{\Gamma(y_i+\frac{1}{\phi})}{\Gamma(y_i+1)\Gamma(\frac{1}{\phi})}\left(\frac{1}{1+\phi\mu_i}\right)^{\frac{1}{\phi}}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i}, \quad (4.7)$$

and

$$\begin{aligned} F(y_i) &= \sum_{s=0}^{y_i}\left[\eta\gamma_0(s) + (1-\eta)\frac{\Gamma(s+\frac{1}{\phi})}{\Gamma(s+1)\Gamma(\frac{1}{\phi})}\left(\frac{1}{1+\phi\mu_i}\right)^{\frac{1}{\phi}}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{s}\right] \\ &= \eta + (1-\eta)\sum_{s=0}^{y_i}\left[\frac{\Gamma(s+\frac{1}{\phi})}{\Gamma(s+1)\Gamma(\frac{1}{\phi})}\left(\frac{1}{1+\phi\mu_i}\right)^{\frac{1}{\phi}}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{s}\right]. \quad (4.8) \end{aligned}$$

5. METHODOLOGY

In this section we consider the implementation of the models discussed in Chapter 4.

5.1 Setup of Variables

The response variable was Y = number of sessions in first EoC - 1 so that standard count data models could be used and we could model zero-inflation rather than one-inflation. Also, for our analysis we only considered the first EoC of each patient during the study period.

Since the EOC data had categorical (nonquantitative) predictor variables (gender, therapist type, procedure type, region), we used *indicator coding* to formulate the model so that these variables would have interpretable coefficients. Indicator coding assigns values 1 and 0 to reflect the presence and absence, respectively, of a level. If a categorical predictor has k levels, it is represented by $k-1$ indicator variables. The level not represented by an indicator variable is called the reference level. For our analysis, we used the most frequent level as the reference level (Table 5.1). For example, because therapist type has three levels (psychiatrist, psychologist, and social worker), we used two indicator variables so each level is uniquely defined by combining the two variables. The coefficient of each indicator variable compares the effect of the specified level to that of the reference level.

Table 5.1: Reference level for each of the categorical variables in the EoC data.

Variable	Reference
Gender	female
Therapist Type	social worker
Procedure Type	individual
Region	south

To minimize the effect of possible high correlation in the parameter estimates,

we centered the age variable. This was done by subtracting the overall age mean of 34.142 from each patient's age.

5.2 Model Fitting

The overall strategy of the Bayesian modeling was to fit the ZICP, ZICGP, and ZICNB models using MCMC. The ZICGP and ZICNB models allowed for zero-inflation, overdispersion, and censoring, while ZICP allowed for zero-inflation and censoring only.

To validate our treatment of censoring, we used Bayesian MCMC to also fit a zero-inflated Poisson (ZIP), a zero-inflated generalized Poisson (ZIGP), and a zero-inflated negative binomial (ZINB) to the noncensored EoCs. The likelihoods of the noncensored models involved just the zero-inflated pdfs and not the survival functions.

We used the following as our prior distribution(s)

$$\begin{aligned} \pi(\beta_0), \pi(\beta_1), ..., \pi(\beta_k) &\sim Normal(0, \sigma^2 = 100000) \\ \pi(eta) &\sim Uniform(0, 1), \\ \pi(\alpha) &\sim Gamma(1, 2), \\ \pi(\phi) &\sim Gamma(1, 2). \end{aligned}$$

All of the Bayesian analyses were done in SAS software Version 9.2 using PROC MCMC (Appendix A). We used the built-in Poisson and negative binomial pdf's and cdf's in SAS software Version 9.2 to specify the likelihood functions of ZIP and ZINB and their censored counterparts. However, for ZIGP and ZICGP we had to hard code the likelihood function. We used 150,000 posterior draws (Table 5.2) from each of our models to estimate the posterior distribution. We examined trace plots of the parameters in order to evaluate convergence. We also used 100 random posterior draws, along with noncensored observations, to calculate the Bayesian goodness of fit

Table 5.2: Bayesian MCMC setup

Setting	Number
Burn	10,000
Iteration	150,000
Thinning	1

statistics for each model where the censoring distribution is

$$g(z) \sim \text{discrete uniform}$$

such that $G(z_i) = 1 - \frac{z_i}{t}$. We chose $t = 350$ as the upper bound on the number of sessions in the first EoC for a patient. The DIC statistic for each model was part of the standard output of the PROC MCMC procedure.

For comparison purposes, we used the LIFEREG procedure in SAS software Version 9.2 to fit Weibull models and the GLM procedure in SAS software Version 9.2 to fit Gaussian models using the same predictors as in the models based on count distributions. For the Gaussian models, we used the log of the number of sessions as the response variable. The Gaussian analysis included the censored observations, but treated them as if they were noncensored. The Weibull analysis adjusted for censoring. We used the maximum likelihood estimates from the Weibull and the Gaussian models to compute the goodness of fit statistics based on the outline in Chapter 2 to calculate the Bayesian χ^2, but with just one repetition. For completeness, both the Weibull and the Gaussian models were also fitted using only the noncensored EoCs.

We also generated a posterior predictive distribution using ZICGP posterior draws for EoCs for 40-year-old females from the South region undergoing individual therapy with a social worker. For the same predictor values we generated the sampling distribution for individual predictions based on the Gaussian model. Both predictive distributions were compared to the observed distribution of EoCs for the same predictor values.

6. RESULTS

This chapter presents the results of the Bayesian analyses and non-Bayesian analyses of the EoC data.

6.1 Noncensored Data

The Bayesian analyses via PROC MCMC ran well for the noncensored data, but took some time (Table 6.1). This was not surprising since we were dealing with a large data set and complex models. After a burn-in of 10,000 iterations, 150,000 iterations were monitored for model analysis. Trace plots for all the parameters in the different models showed adequate convergence (Appendix B).

For analysis involving only the noncensored data, ZIGP had the lowest DIC statistic (Table 6.2). The ZIGP model also had the lowest goodness of fit statistic (3007) even though it was clearly significant. The estimated regression coefficients of ZIGP and ZINB were very similar (Figure 6.3). The ZIP model did not fit the data well.

Table 6.1: Time to run the Bayesian models.

Model	Time (hrs)
ZIP	18
ZINB	21
ZIGP	56
ZICP	19
ZICNB	24
ZICGP	74

The Gaussian model failed to properly account for the zero-inflation present in the data (Figure 6.1), but produced similar parameter estimates to those associated

with the ZIGP and ZINB models. The Weibull model also provided similar parameter estimates.

For all models, the coefficient of procedure type family was significantly negative (Table 6.2). This meant that those undergoing family therapy had fewer sessions on average than those undergoing individual therapy. However, those undergoing both family and individual therapy had more sessions on average than those undergoing individual therapy. It is not clear what these results imply regarding medical offset of family therapy.

6.2 Full Data

As with the noncensored data analysis, after a burn-in of 10,000 iterations, 150,000 iterations were monitored for model analysis. Trace plots for all the parameters in the different models showed adequate convergence. The parameter estimates of all the models that accommodated censoring (all but the Gaussian) were comparable to the estimates based on only noncensored data (Table 6.2 and Table 6.3). Thus censoring appeared to be dealt with correctly.

The estimated intercepts tell an interesting story. The intercepts for those models that accommodated censoring increased between the noncensored data analysis and the full data analysis. This probably is due to the bias of ignoring individuals with large session counts when analyzing the noncensored data. On the other hand, the Gaussian analysis ignored censoring in the full data analysis. Thus the intercept was biased downward by assuming that the censored counts were complete counts.

Of the censored Poisson models (ZICP, ZICGP, and ZICNB), ZICNB had the lowest DIC (Table 6.3). The ZICP model did not fit the data well. For both the ZICGP and ZICNB models, the dispersion parameter was greater than zero and was significant. For the ZICNB model, zero-inflation was not significant while for the ZICGP model zero-inflation was significant. While the parameter estimates were

similar for the ZICGP, ZICNB, Weibull, and even the Gaussian models, properties of predictions based on the models would be very different. As an example, we compared the posterior predictive distribution based on the ZICGP model to the sampling distribution of predictions based on the Gaussian model (back-transformed from the log scale) (Figure 6.2) for 40-year-old females in the South undergoing individual therapy with a social worker. These distributions were very different from each other, although the posterior predictive distribution for the ZICGP model was similar to the empirical distribution of sessions for this subgroup.

Both the ZICGP and ZICNB models had similar parameter estimates (Figure 6.4). However, the goodness of fit statistic suggested that ZICGP (1696) provided a better fit to the full EoC data than ZICNB (5378).

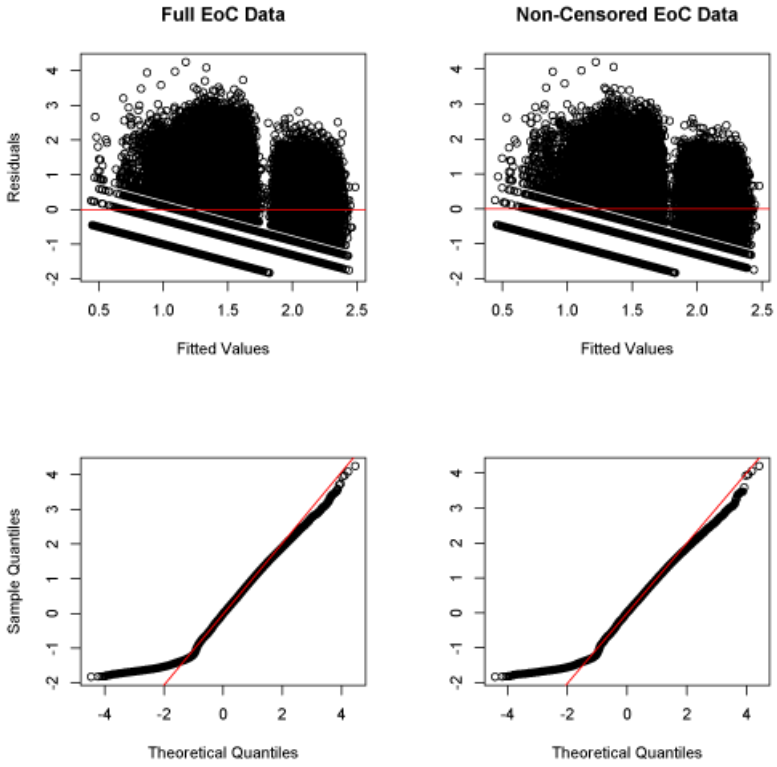

Figure 6.1: Residual plots for the Gaussian models. The top two plots show that the Gaussian model does not fit the data well. The bottom two qq-plots show the lack of fit in the tails, especially the lower tail, signifying that Gaussian models do not properly handle the short left tail of the session count distribution.

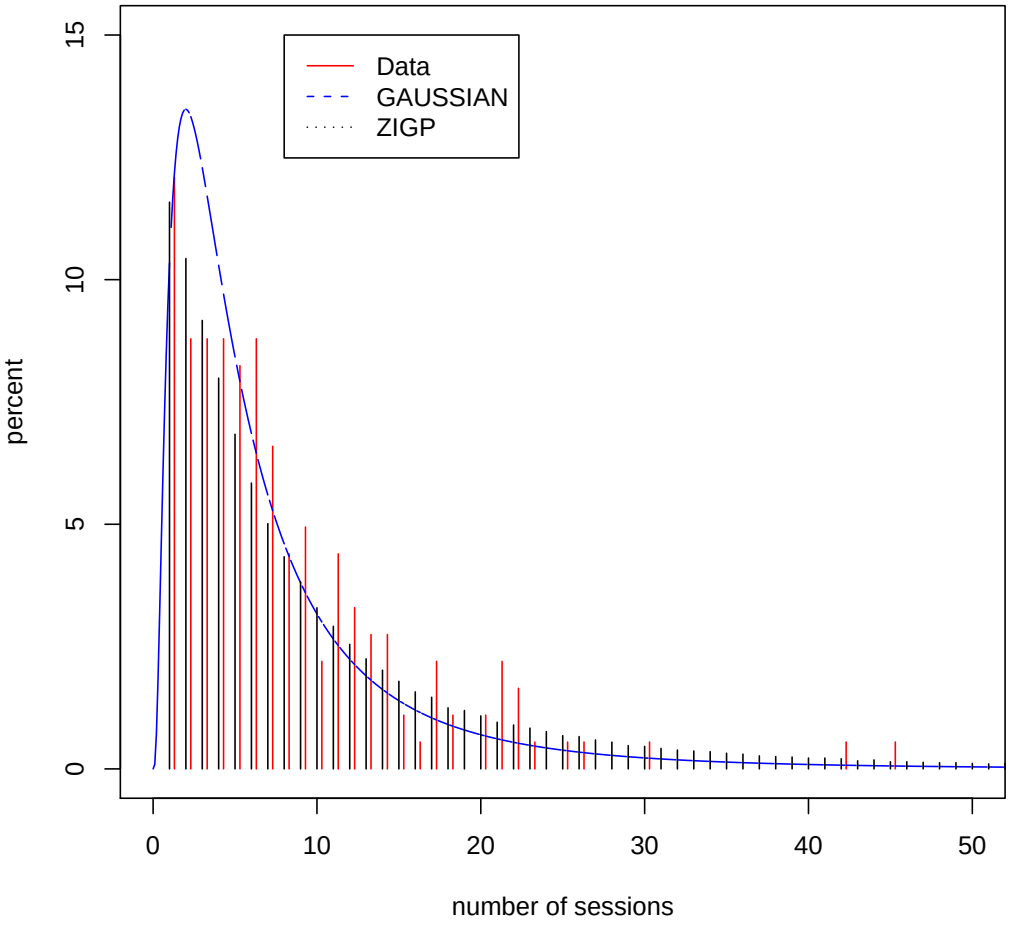

Figure 6.2: Posterior predictive distribution of ZICGP model, sampling distribution for the Gaussian model, and empirical distribution of number of sessions for 40-year-old females in the South undergoing individual therapy with a social worker.

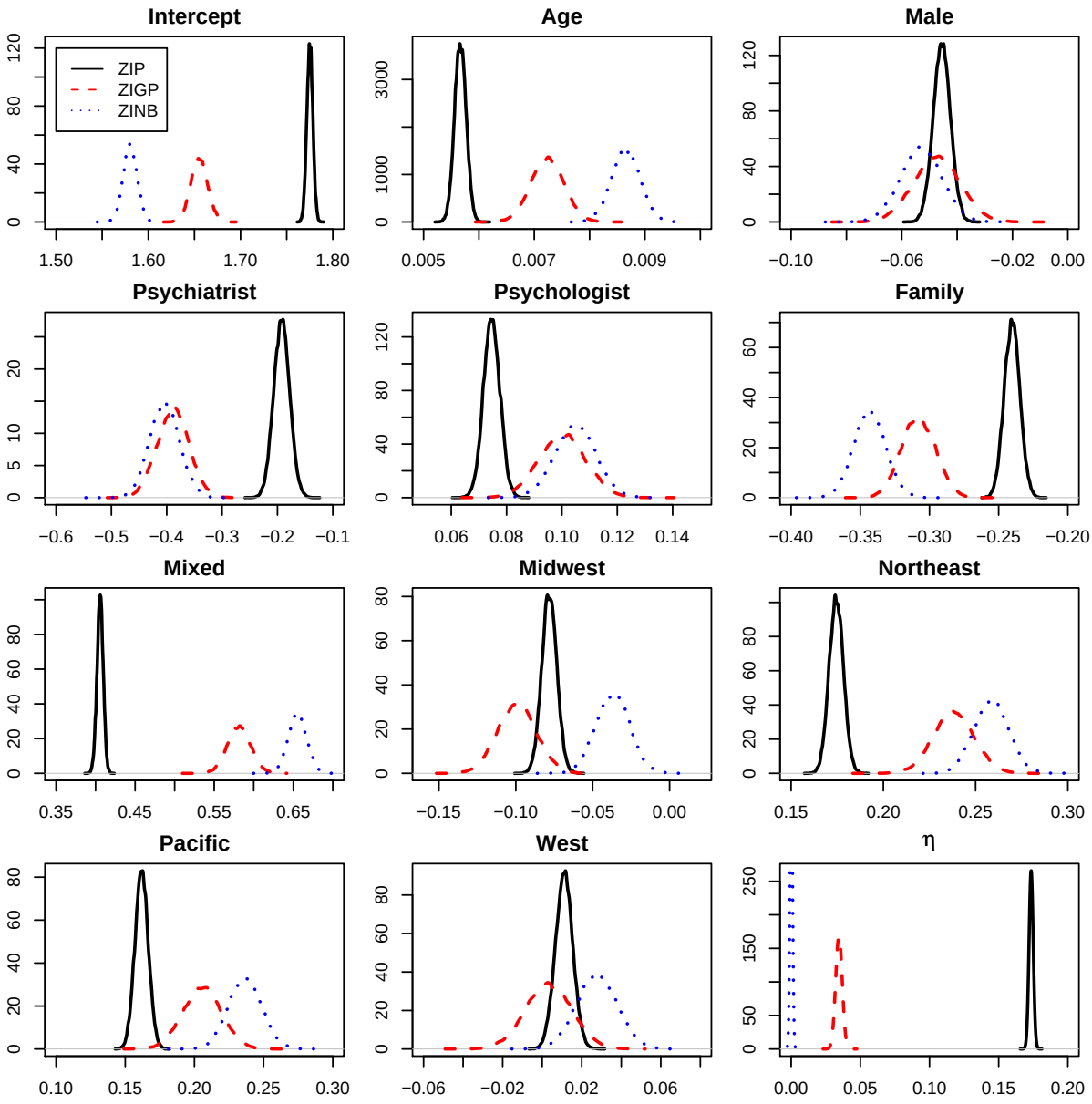

Figure 6.3: Posterior densities for regression coefficients of the ZIP (black solid line), ZIGP (red broken lines) and ZINB (blue dots) models. Overall, the ZIP model underestimated the standard error.

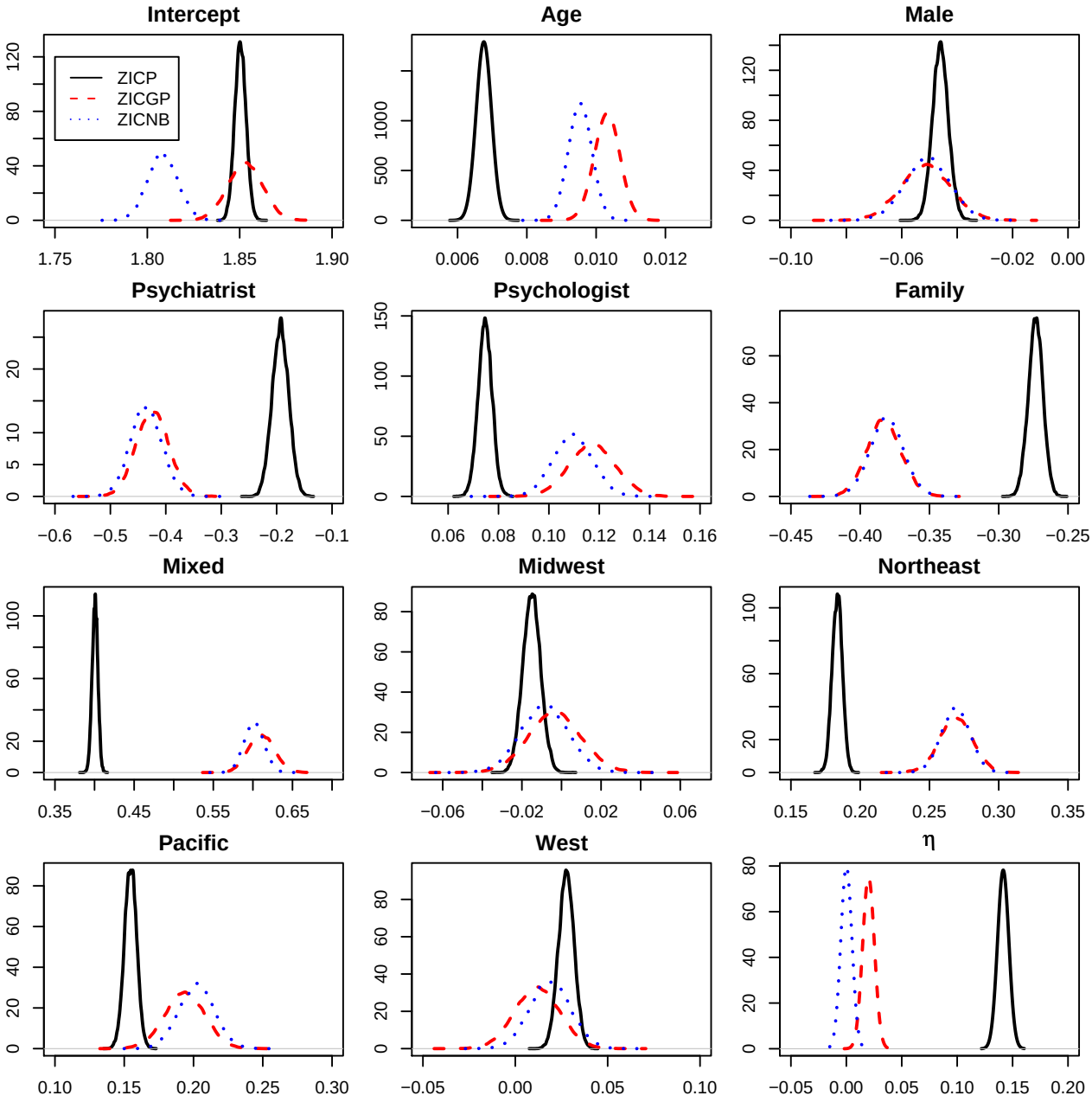

Figure 6.4: Posterior densities for regression coefficients of the ZCIP (black solid line), ZICGP (red broken lines) and ZICNB (blue dots) models. Overall, the ZICP model underestimated the standard error.

Table 6.2: Parameter estimates and standard errors (in parentheses). Only noncensored data were used.

Parameters	Gaussian	Weibull	ZIP[1]	ZIGP[2]	ZINB[3]
Intercept	1.33198	1.83114	1.77510	1.65524	1.580268
	(0.006528)	(0.006592)	(0.003270)	(0.009110)	(0.007565)
Age	0.00456	0.00605	0.00567	0.00724	0.008653
	(0.000215)	(0.000220)	(0.000105)	(0.000300)	(0.000259)
Gender: M	-0.02330	-0.03989	-0.04553	-0.04735	-0.052935
	(0.006371)	(0.006318)	(0.003110)	(0.008464)	(0.007513)
Therapist Type					
Psychiatrist	-0.46968	-0.31334	-0.19232	-0.38866	-0.404091
	(0.022219)	(0.021987)	(0.014463)	(0.028756)	(0.027058)
Psychologist	0.07086	0.08291	0.07462	0.10010	0.104894
	(0.006314)	(0.006268)	(0.002993)	(0.008496)	(0.007345)
Procedure Type					
Family	-0.18701	-0.25892	-0.24030	-0.30798	-0.343256
	(0.009479)	(0.009447)	(0.005635)	(0.012244)	(0.011503)
Mixed	0.70181	0.52292	0.40575	0.58151	0.655238
	(0.010182)	(0.010117)	(0.003967)	(0.014770)	(0.011712)
Region					
Midwest	-0.06587	-0.08304	-0.07819	-0.09925	-0.036996
	(0.009505)	(0.009421)	(0.004899)	(0.012617)	(0.010985)
Northeast	0.21582	0.21084	0.17451	0.23810	0.258471
	(0.008159)	(0.008117)	(0.003897)	(0.011069)	(0.009335)
Pacific	0.18506	0.18350	0.16186	0.20569	0.236692
	(0.010208)	(0.010153)	(0.004792)	(0.013901)	(0.012049)
West	0.00318	0.01096	0.01108	0.00343	0.027661
	(0.008792)	(0.008714)	(0.004401)	(0.011728)	(0.010298)
Other Parameters					
η	-	-	0.17347	0.03452	0.000085
	-	-	(0.001208)	(0.002182)	(0.000085)
α/ϕ	-	-	-	0.32866	1.269804
	-	-	-	(0.002290)	(0.005994)
Fit Statistics					
GoF[4]	13582	22104	58234	3007	5573
DIC[5]	-	-	825746.5	562338.2	686606.6

[1] Zero-Inflated Poisson
[2] Zero-Inflated Generalized Poisson
[3] Zero-Inflated Negative Binomial
[4] Goodness of Fit
[5] Deviance Information Criterion

Table 6.3: Parameter estimates and standard errors (in parentheses). Both censored and noncensored data were used. Adjustments for censoring were made for all the models except the Gaussian.

Parameters	Gaussian[6]	Weibull	ZICP[1]	ZICGP[2]	ZICNB[3]
Intercept	1.29188	1.99994	1.85039	1.85297	1.80844
	(0.005968)	(0.006610)	(0.003050)	0.009417	0.008121
Age	0.00539	0.00802	0.00676	0.01032	0.00954
	(0.000197)	(0.000221)	(0.000098)	(0.000314)	(0.000274)
Gender: M	-0.02425	-0.04241	-0.04603	-0.05137	-0.05079
	(0.005821)	(0.006379)	(0.002881)	(0.009040)	(0.008013)
Therapist Type					
Psychiatrist	-0.44879	-0.33507	-0.19314	-0.42412	-0.43578
	(0.020472)	(0.02219)	(0.014928)	(0.029346)	(0.028449)
Psychologist	0.06733	0.09349	0.07467	0.11712	0.10956
	(0.005762)	(0.006326)	(0.002780)	(0.009107)	(0.007808)
Procedure Type					
Family	-0.20559	-0.31193	-0.27338	-0.38280	-0.38083
	(0.008755)	(0.009540)	(0.005289)	(0.012464)	(0.012207)
Mixed	0.73291	0.51109	0.40055	0.61084	0.60139
	(0.009496)	(0.010211)	(0.003688)	(0.016469)	(0.012481)
Region					
Midwest	-0.00446	-0.01170	-0.01505	-0.00337	-0.00869
	(0.008622)	(0.009516)	(0.004457)	(0.013376)	(0.011979)
Northeast	0.20909	0.22783	0.18346	0.26924	0.268861
	(0.007437)	(0.008193)	(0.003642)	(0.011934)	(0.010156)
Pacific	0.19115	0.16999	0.15457	0.19378	0.202161
	(0.009415)	(0.010249)	(0.004397)	(0.014508)	(0.012824)
West	0.01627	0.01843	0.02761	0.01198	0.018939
	(0.008072)	(0.008795)	(0.004200)	(0.012196)	(0.011028)
Other Parameters					
η	-	-	0.14154	0.02002	0.000056
	-	-	(0.001010)	(0.001775)	(0.000057)
α/ϕ	-	-	-	0.32681	1.198601
	-	-	-	(0.002046)	(0.005958)
Fit Statistics					
GoF[4]	21440	20269	63583	1696	5378
DIC[5]	-	-	920194.2	598571.7	597665.0

[1] Zero-Inflated Poisson
[2] Zero-Inflated Generalized Poisson
[3] Zero-Inflated Negative Binomial
[4] Goodness of Fit
[5] Deviance Information Criterion
[6] Censoring was ignored

7. CONCLUSIONS

We were able to successfully fit the zero-inflated, overdispersed and censored count models to the EoC data using the PROC MCMC procedure in SAS software Version 9.2. The code used can be found in Appendix C. We modified the Bayesian goodness of fit procedure to apply to censored EoC count. This was original work not found in the literature on zero-inflated models. For the goodness of fit statistic, we assumed that the censoring distribution was uniform. The code used for the analysis can be found in Appendix D.

The positive significant estimates of the overdispersion parameters ϕ and α meant that the EoC data had overdispersion. Thus, ZIP and ZICP models did not fit the data well at all. Also, the ZIP and ZICP models seriously underestimated standard errors (Figures 6.3 and 6.4), which present the issue of biases in the model. The best fitting model for the EoC data was the ZICGP model. It had the second lowest DIC value (very close to that of ZICNB) and had the best goodness of fit statistic. Unfortunately, the ZICGP model still had significant lack of fit. The model might be improved by including interactions between predictor variables and squares of predictor variables in the model. Another option would be to include random effects in the model for the mean of the Poisson distribution, modelling the mixing proportion as a logistic function of predictors, or modelling the count distribution as a mixture of two Poissons plus a point mass at zero.

A comparison of the ZICGP and the ZIGP models for noncensored data confirmed that our procedures were properly handling the censoring. Even though the ZICGP model had significant lack of fit, it still fitted the EoC data much better than any other model, especially the Gaussian and the Weibull models. The Gaussian models, which ignored censoring, resulted in biased conclusions.

Also, with the ZINB and the ZICNB models, the zero-inflation parameter η was not significant in fitting the EoC data. This means that the NB distibution does well in accommodating zero-inflation in the EoC data, unlike the Poisson distribution.

Inaccurate and misleading prediction intervals were obtained by using a Gaussian (continuous distribution) as a model for the EoC data. However, inferences about the regression parameters were fairly good for the Weibull model. Presumably these inferences would also be good for a censored Gaussian model. When prediction is the goal, as it might be for an insurance company that used these data to predict the number of sessions for a client, the ZICGP model must be used. The Gaussian model clearly under-predicts the number of sessions (Figure 6.2), which may lead to an inaccurate prediction of cost for the insurance company.

All in all, we were successfully able to fit the complex models to the EoC data, and in doing so prove the need of addressing the issues of zero-inflation, overdispersion, censoring, and goodness of fit for censored count data.

BIBLIOGRAPHY

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1990), *Statistical Modeling in GLIM*, Oxford University Press.

Arjas, E. and Gasbarra, D. (1994), "Nonparametric Bayesian Inference From Right Censored Survival Data, Using The Gibbs Sampler," *Statistica Sinica*, 4, 505–524.

Bohning, D. (1994), "A Note on Test for Poisson Over Dispersion," *Biometrika*, 81, 1–14.

Cameron, A. C. and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge, U.K.: Cambridge University Press.

Cao, J., Moosman, A., and Johnson, V. (2007), "A Bayesian Chi-Squared Goodness-of-Fit Test for Censored Data Models," *Biostat*, 43, Berkeley Electronic Press.

Caudill, S. and Mixon, F. (1995), "Modeling Household Fertility Decisions: Estimation and Testing Censored Regression Models For Count Data," *Empirical Economics*, 20, 183–196.

Crane, D. R. and Christenson, J. D. (2007), "The Medical Offset Effect: Patterns in Outpatient Services Reduction for High Utilizers of Health Care," *Contemporary Family Therapy*, 2008, 127–138.

El-Sayyad, G. (1973), "Bayesian and Classical Analysis of Poisson Regression," *Royal Statistical Society*, 35, 445–451.

Famoye, F. and Singh, K. P. (2006), "Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data," *Data Science*, 4, 117–130.

Famoye, F. and Wang, W. (2004), "Censored Generalised Poisson Regression Model," *Computational Statistics and Data Analysis*, 46, 547–560.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Boca Raton, FL: CRC Press., 2nd ed.

Greene, W. (1994), *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models.*

Gurmu, S. and Trivedi, P. (1996), "Excess Zeros in Count Models for Recreational Trips," *Business and Economic Statistics*, 14, 469–477.

Hall, D. (2000), "Zero-inflated Poisson and Binomial Regression with Random Effects: A Case Study," *Biometrics*, 56, 1030–1039.

Ismail, N. and Jemain, A. A. (2007), "Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models," *Casualty Actuarial Society Forum*, 2007, Winter.

Johnson, V. E. (2004), "A Bayesian χ^2 test for Goodness-of-Fit," *Annals of Statistics*, 32, 2361–2384.

Lambert, D. (1992), "Zero-Inflated Poisson Regression Models with an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14.

Lo, A. Y. (1992), "Bayesian Inference For Poisson Process Models With Censored Data," *Nonparametric Statistics*, 2, 71–80.

Morgan and Scovial (2007), "Medical Offset in Health Care Data," , Brigham young University.

Renshaw, A. (1994), "Modeling the Claims Process in the Presence of Covariates," *ASTIN Bulktin*, 24, 265–285.

Rideout, M., Hinde, J., and Demetrio, C. G. B. (2001), "A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives," *Biometrics*, 57, 219–223.

Rodrigues-Motta, M., Gianola, D., Heringstad, B., Rosa, G., and Chang, Y. (2006), "Zero-Inflated Poisson Model for Genetic Analysis of the Number of Mastitis Cases in Norwegian Red Cows," *Dairy Science*, 90, 5306–5315.

Selvin, S. (1974), "Maximum Likelihood Estimation in the Truncated or Censored Poisson Distribution," *American Statistical Association*, 69, 234–237.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit," *J. R. Staistics Society*, 64, 583–639.

Terza, J. (1985), "A Tobit-Type Estimator for the Censored Poisson Regression Model," *Econom. Lett.*, 18, 361–365.

van den Broek, J. (1995), "A Score Test for Zero-inflation in a Poisson Distribution," *Biometrics*, 51, 738–743.

Winkelmann, R. and Zimmermann, K. F. (1995), "Recent Developments in Count Data Modeling: Theory and Application," *Economic Survey*, 9, 1–4.

Yin, G. (2009), "Bayesian goodness-of-fit test for censored data," *Statistical Planning and Inference*, 139, 1474–1483.

Zorn, C. (1996), "Evaluating Zero-inflated and Hurdle Poisson Specifications," *Midwest Political Science Asociation*, April 18-20, 1–16.

A. APPENDIX

PROC MCMC Implementation

PROC MCMC in SAS software Version 9.2 is a flexible simulation-based procedure that is suitable for fitting a wide range of Bayesian models (SAS/STAT 9.2 User's Guide, The MCMC Procedure) . PROC MCMC obtains samples from the corresponding posterior distributions, produces summary and diagnostic statistics, and saves the posterior samples in an output data set that can be used for further analysis. One can analyze data that have any likelihood function(s) or prior(s) with PROC MCMC, as long as these functions are programmable using the SAS DATA step functions. The parameters can enter the model linearly or in any nonlinear functional form. The default algorithm that PROC MCMC uses is an adaptive blocked random walk Metropolis algorithm that uses a normal proposal distribution. For our models, we used the following input commands in PROC MCMC:

- **DATA**= names the input data set
- **OUTPOST**= names the output data set for posterior samples of parameters. If no OUTPOST command is specified then the posterior draws are not available for any further analysis such as goodness of fit statistics.
- **NBI**= 1,000, specifies 1,000 burn-in iterations.
- **NMC**= 150,000, specifies 150,000 MCMC iterations, excluding the burn-in iterations
- **THIN**= 1, specifies the thinning rate of 1. A thinning of THIN= n, means to keep every nth simulation sample and discard the rest. All of the posterior statistics and diagnostics are calculated using the thinned samples. With

(THIN=1) we kept all the simulation draws. Increasing the thinning rate helps decrease the auto-correlation in the parameter values. A high auto-correlation in the parameter values deflate's the precision of the estimates.

- **PROPCOV**= QUANNEW, specifies the method used in constructing the initial covariance matrix for the Metropolis-Hastings algorithm. For our analysis we used the Quasi-Newton (QUANNEW) method.
- **PRIOR**, assign prior distribution for the parameters with PRIOR statement.
- **DIC**, computes the deviance information criterion (DIC) for the model.

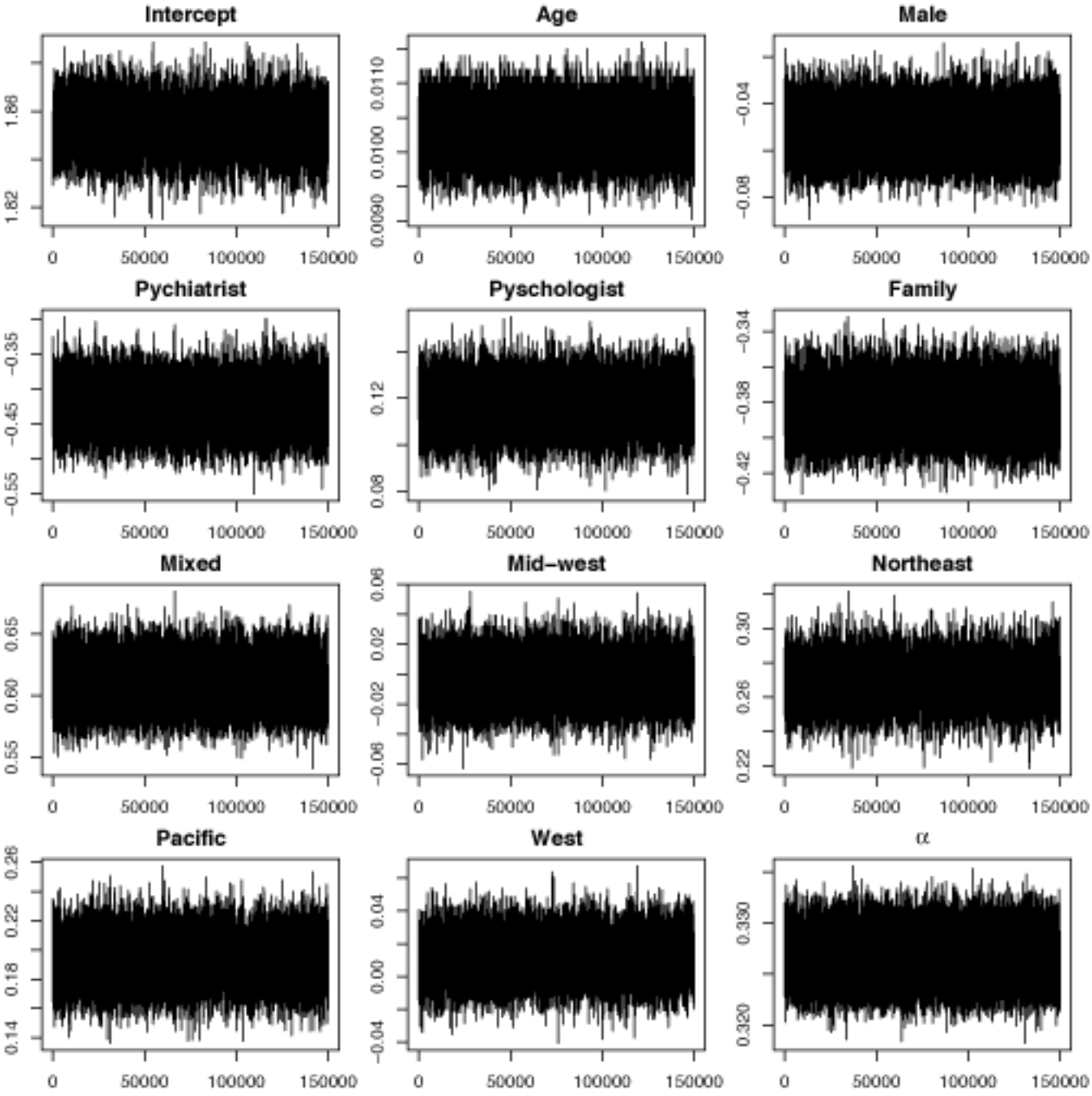

Figure B.1: Mixing plots of the ZICGP parameters. All the parameters show adequate convergence.

C. APPENDIX

Bayesian MCMC Code for ZICGP

Below is the Baysian MCMC analysis code for the ZICGP model using SAS software Version 9.2. The ZICGP model required hard coding of the likelihood function.

```
*/read in the EoC data set (already with indicators);
data d1;
infile ".../ZICGP/EOC_DATA_MCMC.csv"
 firstobs=2  dlm=',';
input session censored intercept age female male psychiatrist
 psychologist socialworker
 individual family mixed west midwest pacific south northeast;
ses=session - 1; */adding a column for (session-1) to make
 session zero-inflated;
agec= age - 36.41745;
run;
proc print data=d1 (obs=10); */looking at data;
run;


*Baysian MCMC for Zero-Inflated Censored Generalized Poisson (ZICGP) model;
ods graphics on;
proc mcmc DATA=d1 SEED=2009 NBI=10000 NMC=150000  NTU=1000
 outpost=postZICGP  propcov=quanew DIC monitor =(_parms_  mu);
ods select Parameters PostSummaries PostIntervals DIC ess mcse tadpanel;
parms eta 0.142 alpha 1;
```

```
parms beta0 1.74 beta1 0.0068 beta2 0.05;
parms beta3 -0.193 beta4 0.08 beta5 -0.27 beta6 0.406;
parms beta7 -0.0191 beta8 -0.147 beta9 -0.189 beta10 -0.1747;
prior beta0 beta1  beta2  beta3  beta4  beta5  beta6  beta7
 beta8 beta9   beta10 ~ normal(0,var=1000000);
prior eta ~ uniform(0,1);
prior alpha ~ gamma(1,scale=2);
mu=exp(beta0*intercept +  beta1*agec  + beta2*male +
 beta3*psychiatrist  + beta4*psychologist  + beta5*family
 + beta6*mixed  + beta7*midwest +
 beta8*northeast + beta9*pacific + beta10*west);
a=(mu/(1+alpha*mu));
cdf=0;
if (ses >0) then do i =0 to (ses-1);
val= pdf("Poisson",i,a)*((1+alpha*i)**(i-1))*
 exp(-mu*alpha*i/(1+alpha*mu));
cdf=cdf + val;
end;
else cdf= pdf("Poisson",ses,a)*((1+alpha*ses)**(ses-1))*
 exp(-mu*alpha*ses/(1+alpha*mu));
if censored = 0 then do;
like=(eta*(ses eq 0)  +
(1-eta)*pdf("poisson",ses,a)*((1+alpha*ses)**(ses-1))*
 exp(-mu*alpha*ses/(1+alpha*mu))); if like <= .000000000000001
 then like=.000000000000001;
llike=log(like);end;
else  do; like=(1 - eta)*(1 - cdf);  if like <= .000000000000001
```

```
 then like=.000000000000001;
llike=log(like);end;
model general(llike);
run;
ods graphics off;

*/Exporting the ZICGP posterior information to file;
proc export data=postZICGP outfile='.../ZICGP/postZICGP.csv';
run;
```

D. APPENDIX

Goodness of Fit Analysis Code

Below is the Bayesian χ^2 goodness of fit analysis code for ZICGP using SAS software Version 9.2.

```
*/read in the EoC data set (already with indicators);
data d1;
infile ".../EOC_DATA_MCMC.csv"  firstobs=2  dlm=',';
input session censored intercept age female male psychiatrist
 psychologist socialworker individual family mixed west midwest
 pacific south northeast;
ses=session - 1; */adding a column for (session-1) to make
 session zero-inflated;
agec= age - 36.41745;
run;

*Bayesian Chi Squared;
data postZICGP;
infile "...ZICGP/postZICGP.csv"  firstobs=2  dlm=',';
input Iteration eta alpha beta0 beta1 beta2 beta3 beta4 beta5
 beta6 beta7 beta8 beta9 beta10 mu LogPrior LogLike LogPost;
dumb=ranuni(0);
proc sort data=postZICGP;
by dumb;
run;
```

```
proc iml;
use postZICGP;
read all var{beta0 beta1 beta2 beta3 beta4 beta5 beta6 beta7
 beta8 beta9 beta10 eta alpha} into postd;
use d2;
read all var{intercept agec male psychiatrist psychologist family
 mixed  midwest northeast pacific west ses} into dat;
n=nrow(dat);
m=nrow(postd);
irtp=pdf("NORMAL",0); * 'inverse root two pi'  for stirlings formula;
q=j(300,1,0);
preds=j(m,1,0);
do j=1 to n;
  mu= exp(postd[j,1:11]*t(dat[j,1:11]));
  alpha=postd[j,13];
  a=(mu/(1+alpha*mu));
  do i=1 to 300;
    ii=i-1;
    if ii <= 30 then q[i,1]= (postd[j,12]*(ii=0)+ (1-postd[j,12])
    *exp(ii*log(a)-a-log(gamma(ii+1))+(ii-1)*log(1+alpha*ii) - a*alpha*ii));
    else
    q[i,1]= (postd[j,12]*(ii=0)+ (1-postd[j,12])*irtp*exp(ii*log
    (a)-a-(ii+.5)*log(ii)+ii+(ii-1)*log(1+alpha*ii) - a*alpha*ii));
  end;
  call randgen(x,'TABLE',q);
  preds[j,1]=x;
end;
```

```
do j=n+1 to m;
  mu= exp(postd[j,1:11]*t(dat[n,1:11]));
  alpha=postd[j,13];
  a=(mu/(1+alpha*mu));
  do i=1 to 300;
    ii=i-1;
    if ii <= 30 then q[i,1]= (postd[j,12]*(ii=0)+ (1-postd[j,12])*
   exp(ii*log(a)-a-log(gamma(ii+1))+(ii-1)*log(1+alpha*ii) - a*alpha*ii));
    else
    q[i,1]= (postd[j,12]*(ii=0)+ (1-postd[j,12])*irtp*exp(ii*log(a)
    -a-(ii+.5)*log(ii)+ii+(ii-1)*log(1+alpha*ii) - a*alpha*ii));
  end;
  call randgen(x,'TABLE',q);
  preds[j,1]=x;
end;
preds=((dat[,12]+j(n,1,1))//j(m-n,1,-9))||preds;
*print preds;
create pred from preds;
append from preds;
proc print data=pred (obs=100);
run;
```