

Fall 12-2018

CENSORED COUNT DATA ANALYSIS – STATISTICAL TECHNIQUES AND APPLICATIONS

Xiao Yu
UTHealth SPH

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen



Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

Recommended Citation

Yu, Xiao, "CENSORED COUNT DATA ANALYSIS – STATISTICAL TECHNIQUES AND APPLICATIONS" (2018). *UT School of Public Health Dissertations (Open Access)*. 1.
https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/1

This is brought to you for free and open access by the School of Public Health at DigitalCommons@TMC. It has been accepted for inclusion in UT School of Public Health Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

CENSORED COUNT DATA ANALYSIS – STATISTICAL TECHNIQUES
AND APPLICATIONS

by

XIAO YU, BS, MS

APPROVED:

WENYAW CHAN, PHD

LUNG-CHANG CHIEN, DRPH

JOHN M. SWINT, PHD

KAI ZHANG, PHD

DEAN, THE UNIVERSITY OF TEXAS

SCHOOL OF PUBLIC HEALTH

Copyright

by

Xiao Yu, BS, MS, PHD

2018

CENSORED COUNT DATA ANALYSIS – STATISTICAL TECHNIQUES
AND APPLICATIONS

by

XIAO YU

BS, Shaanxi University of Science and Technology, 2011

MS, University of Illinois at Urbana - Champaign, 2013

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas

December, 2018

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my academic advisor Dr. Wenyaw Chan for his unwavering support throughout my PhD program. Dr. Chan provided me excellent advice throughout the course of my dissertation.

I am very grateful to my dissertation supervisor, Dr. Lung-Chang Chien for his excellent guidance on my dissertation project. Dr. Chien came up with the idea of conducting the analysis on censored count data, which was commonly used in spatial analysis and an ideal project for my dissertation.

I am also grateful to all my committee members Dr. John M. Swint, Dr. Kai Zhang and Dr. Folefac D. Atem for all the support they have provided to help me complete my dissertation successfully.

I would also like to express my deepest gratitude to all my family and friends. I have them by my side through both good and bad times and I would not have been possible to finish my dissertation without their strong support.

CENSORED COUNT DATA ANALYSIS – STATISTICAL TECHNIQUES AND APPLICATIONS

Xiao Yu, BS, MS, PhD

The University of Texas

School of Public Health, 2018

Dissertation Chair: Wenyaw Chan, PhD

Count data are commonly used to report frequency statistics of diverse health outcomes. However, some data are marked intentionally to avoid leaking information that could be used to identify individuals when population sizes are small. The situation hinders the further use from those data in public health research. Thus, an accurate and efficient method for dealing with censored count data is needed.

We developed Integrated Nested Laplace Approximation algorithm to censored Poisson regression model to deal with censored count data and improve the computational efficiency. In addition, we applied three methods to deal with censored count data: 1) multiple imputation (MI); 2) small area estimation (SAE); 3) censored Poisson regression model (CPRM) and compared the accuracy and efficiency of these three methods.

A series of simulations results in that the censored Poisson regression method conducted the closest estimates to the true values (with the relative error = 0.21%), and MI had the worst results (with relative error=9.13%) under the censored proportion by 7.9 %.

After comparing the results under the censored proportion by 33.61% and 54.1%, the censored Poisson regression method still showed a smaller relative error than the other two methods.

We also applied these three methods to assess the association between heat wave temperature and hospitalization due to cardiovascular diseases in Harris County, Texas, from 2006 to 2011. By comparing the relative errors and bar plots across different methods under different censored proportions, we concluded that by considering the balance of the estimation accuracy with computational time, the censored Poisson regression model is the best method for dealing with censored count datasets under different censored proportions, especially when the censored proportions were less than 30%.

TABLE OF CONTENTS

| | |
|--|-----|
| List of Tables | i |
| List of Figures | ii |
| List of Appendices | iii |
| BACKGROUND | 1 |
| Literature Review | 1 |
| Public Health Significance | 8 |
| Specific Aims | 8 |
| METHODS | 10 |
| Methods for Specific Aim 1 | 10 |
| Methods for specific Aim 2..... | 11 |
| Method for Specific Aim 3..... | 13 |
| Simulation Study | 15 |
| Simulation Procedure | 15 |
| Application of Multiple Imputation to Simulated Censored Count Datasets | 17 |
| Application of Small Area Estimation to Simulated Censored Count Datasets..... | 17 |
| Application of Censored Poisson Regression to Censored Count Dataset..... | 19 |
| Case study | 20 |
| Data Description | 20 |
| Data Analysis..... | 22 |
| Limitations | 23 |
| Strengths..... | 23 |
| APPENDICES | 25 |
| Appendix A | 25 |
| REFERENCES | 26 |
| Journal Article #1 | 31 |
| Development of Integrated Laplace Approximation in the Censored Poisson Regression Model | 31 |
| Journal Article #2 | 50 |
| Performance of Multiple Imputation, Small Area Estimation, and Censored Poisson Regression Model When Handling Censored Count Data..... | 50 |

LIST OF TABLES

| | |
|--|----|
| Tables for Journal Article #1 | 46 |
| Table 1. Simulation results for censored Poisson regression model under different censored proportions. | 46 |
| Table 2. Case study results for censored Poisson regression model under different censored proportions. | 46 |
| Table 3. Case study results for censored Poisson regression model under different censored proportions | 46 |
| Tables for Journal Article #2 | 69 |
| Table 1. Parameter estimation results across three methods under different censored proportions | 69 |
| Table 2. Average relative errors for different methods under different censored proportions | 69 |
| Table 3. Average relative errors for all methods under different censored proportions | 69 |
| Appendix Table 1. Case study results under censored point = 7 (censored proportion = 4.56%) | 81 |
| Appendix Table 2. Case study results under censored point = 8 (censored proportion = 8.08%) | 81 |
| Appendix Table 3. Case study results under censored point = 9 (censored proportion = 12.96%) | 82 |
| Appendix Table 4. Case study results under censored point = 10 (censored proportion = 20.41%) | 82 |
| Appendix Table 5. Case study results under censored point = 11 (censored proportion = 28.07%) | 82 |
| Appendix Table 6. Case study Results under censored point = 12 (censored proportion = 37.88%) | 83 |
| Appendix Table 7. Case study results under censored point = 13 (censored proportion = 48.65%) | 83 |
| Appendix Table 8. Case study results under censored point = 14 (censored proportion = 59.33%) | 83 |

LIST OF FIGURES

| | |
|---|----|
| Figures for Journal Article #1 | 47 |
| Figure 1. Figure 1. Flow Chart for Simulating Censored Dataset. | 47 |
| Figures for Journal Article #2 | 70 |
| Figure 1. Bar plot for estimations under censored point = 7 (censored proportion = 4.56%) . | 70 |
| Figure 2. Bar plot for estimations under censored point = 8 (censored proportion = 8.08%) . | 71 |
| Figure 3. Bar plot for estimations under censored point = 9 (censored proportion = 12.96%) | 72 |
| Figure 4. Bar plot for estimations under censored point = 10 (censored proportion = 20.41%) | 73 |
| Figure 5. Bar plot for estimations under censored point = 11 (censored proportion = 28.07%) | 74 |
| Figure 6. Bar plot for estimations under censored point = 12 (censored proportion = 37.88%) | 75 |
| Figure 7. Bar plot for estimations under censored point = 13 (censored proportion = 48.65%) | 76 |
| Figure 8. Bar plot for estimations under censored point = 14 (censored proportion = 59.33%) | 77 |
| Figure 9. Bar plot for S.D. of Maximum heat estimates under different censored proportions | 78 |
| Figure 10. Relative errors under different censored proportion across methods | 79 |
| Appedix Figure 1. Bar plot for parameter estimation for different methods under different proportions | 80 |
| Appedix Figure 2. Bar plot for standard deviation for parameter estimation for different methods under different proportions..... | 80 |

LIST OF APPENDICES

| | |
|---------------------------------------|----|
| Appendic for Journal Article #2 | 80 |
|---------------------------------------|----|

BACKGROUND

Literature Review

In public health, count data sets are commonly used to report statistics related to emerging or existing health problems, and play a significant role in biostatistics. In fact, most health reports published by the U.S. Centers for Disease Control and Prevention (CDC) are based on count data. For example, the CDC's Birth Defects Countries and Organizations United for Neural Tube Defects Prevention initiative reports that 3,000 pregnancies in the United States are affected by neural tube defects each year [1] and estimates that folic acid fortification may reduce the prevalence of neural tube defects by 50% or more. In this case, count data sets are being used to help prevent neural tube defects, and associated morbidity and mortality rates. In addition, the CDC reports count data for cases of Lyme disease by county, state, and year, which allows the prevalence of Lyme disease to be analyzed geographically and temporally. Data show that cases of Lyme disease are concentrated in the Northeast and Upper Midwest regions of the United States, which enables targeting of prevention efforts, i.e., those states with a higher prevalence of Lyme disease can dedicate more resources to prevent it [2].

Across public health disciplines, count data are commonly analyzed using Poisson regression models. For example, a study on lung cancer mortality and cigarette smoking used a Poisson regression model to estimate lung cancer deaths among physicians who were regular cigarette smokers [3]. A cervical cancer study used a generalized linear Poisson regression model to assess geographic heterogeneity in human papillomavirus [4]. Loomis, Richardson, and Elliott (2005) used a Poisson regression model to examine the association

between brain cancer and exposure to magnetic fields among a cohort of 138,905 male electrical workers in the United States [5]. A health behavior study used a Poisson regression model to examine the relationship between the number of alcoholic drinks and demographic characteristics [6]. Lastly, a study of illness and injury surveillance used a Poisson regression model to monitor morbidity, and to assess the overall health of the Department of Energy workforce [7]. A major part of the count data are used the Poisson regression model; however, in reality, the Poisson regression model needs more restrictions because of the unique property of Poisson distribution.

Poisson distribution has a very unique property: the mean of the distribution must be equal to the variance of the distribution. If a dataset has a large number of zeros, the over-dispersion problem (the variance larger than the mean) emerges. Therefore, based on the regular Poisson regression model, many other forms related to the Poisson regression model have been derived. For example, a study on death notice data in London used zero-adjusted generalized Poisson model, in which the dataset has more zeros than expected [8]. An occupation injury prevention program used the zero-inflated Poisson regression model with random effects to evaluate the injury. Usually, they used the Poisson regression model to analyze the injury counts; however, in this case, over 65% of the observations are zeros. So, they adopted Newton-Raphson and quasi-Newton algorithms to fit the zero-inflated Poisson regression [9]. A paper systematically introduced the zero-inflated model and zero-truncated model and gave some comparison criterion for each model [10].

Another type of count data is recorded by locations, such as the number of cancer cases in each county; and an advanced Poisson regression model is needed to take spatial

autocorrelation into account. For example, weighted Poisson regression models use a spatial weighting function to estimate spatial variations among Poisson regression parameters. A research on geographical distribution of working-age mortality in the Tokyo metropolitan area used the weighted Poisson regression to analysis the death count in each area [11]. A United Kingdom study used the weighted Poisson regression to model the under-dispersed data of clutch sizes. Although the data related to the weather condition and changes in the geographical locations, the weighted Poisson regression still fitted the data well [12]. (Spatial Poisson regression model overcomes the problem of disparate discretization by relating all spatially varying quantities to a random field model.) An research on the effect of traffic pollution on respiratory disorders in children used the Spatial Poisson regression through Markov chain Monte Carlo method [13].

In public health, one situation commonly exists in the count data, especially for spatial analysis, is the censored situation in the count data sets. Censored data have two fundamental types. One is left censored data, which means that a data point is below a certain value, but it is unknown by how much [14]. For example, in the CDC's Childhood Lead Poisoning Prevention Program surveillance database, the number of children with elevated blood lead levels in each county is censored by 5 or less [15]. Another one is right censored data, which means that an unknown point is somewhere above the certain value. In health behavior study, such as those examining alcohol consumption patterns among male college students, binge drinking may be defined as "five or more drinks in one sitting" and may code the dependent variable as "0," "1," "2," "3," "4," or "5 or more drinks"; values greater than 5 would be censored. In the 2000 US Census individual census report, the question about the

number of persons rode to work in one car, truck or van last week, the possible answers will be 1, 2, 3, 4, 5, 6 and 7+. The numbers larger than 7 are censored [16]. In a research of women's fertility, the numbers of children that a woman has been censored when the number is larger than 5 [16, 17].

Censored count data are formed for several reasons. The major one is that the data are unreleased on purpose, such as to avoid reporting information that can be used to identify individuals when population sizes are small. For example, the Illinois Department of Public Health Sexually Transmitted Diseases dataset does not publish sexually transmitted disease data for counties with a population less than 15,000 or with a total birth rate less than 300, so the null values in the dataset are censored data [18]. Moreover, the survey design might cause censored data. For example, in some survey design, when the possible answer to a questions is 0, 1, 2, 3, and 4+, then the number larger than 4 is the censored part in this situation [19]. The example in 2000 US Census individual census report of the number of persons rode to one car, which we mentioned before, is the same situation [16]. Censored count data are usually considered as missing data, which are initially excluded in data analysis because of the principle of complete case analysis[20]. For example, when Frome (1981) used the Poisson regression model to deal with the count data, the censored part was defined as missing values[3].

In previous studies, censored data were considered as missing, however, censored data are not actually missing, but just intentionally masked, which may lead to biased results. For example, a study applying the Cox regression model on the reduction of blood lead levels shows that the results with the censored data are less biased than those without the censored

data [21, 22]. Even in the simple linear regression model, removing the censored data could cause the bias in the estimation [23]. Furthermore, in a study of hypertension treatment using the generalized hierarchical multivariate conditional autoregressive model, when 24 censored data points were considered missing, 80% of patients completed the protocol with effective control of hypertension and no side effects; however, when the censored data were accounted for, the percentage of patients was 44% instead of 80% [24, 25].

One way to deal with the censored count data is the multiple imputation method. The mission of multiple imputation is to create a complete data set, then statistical models could be used as usual. Imputation method was first developed to deal with missing data problem in the 1980s [26, 27]. Then, the multiple imputation method begun widely used in practical research, including estimating the distribution of time from HIV seroconversion to AIDS, completing the health care survey data, and finishing the patients' information on radiographic measurements to detect whether the prosthesis is loosening [28-30]. With the development of computer and software, the multiple imputation method was then widely used in different software packages and applied to more datasets [31-35]. Zhou et al. (2001) compared multiple imputation methods with the mean imputation method and applied the multiple imputation in public health research data [36]. The multiple imputation method not only use on the outcome variable, but also could be used no the covariates. For example, a blood pressure study applied the multiple imputation method to the covariates of missing values, including the family income and family earnings. A research on the National Health data used multiple imputation to the income variable [37-39]. Nowadays, the multiple

imputation method has been developed to solve advanced missing data problems, such as nonparametric multiple imputation, multilevel multiple imputation, and so on [39-43]

A second way to deal with the censored part of the count data, which is frequently used in the spatial analysis, is the small area estimation method. Small area estimation is used when traditional demographic sample surveys designed for national estimates do not provide large enough samples to produce reliable direct estimates for small areas such as counties. Small area estimation has a long history. It first existed in 11th century England, then many other countries began to have the similar history [44]. All of them focused on demographic methods. Nowadays, the statistical methods using for small area estimation developed dramatically. Bayesian unit-level model estimates the prevalence of diabetes at each county in the U.S. [45]. This model can analyze the prevalence for each county level considering different independent variable layers. For example, the method could estimate the prevalence for the specific county, specific gender and specific age group. Then the model was extended by changing the distribution of the prevalence from Poisson to Binomial, and estimated the diabetes incidence for each county by different layers [46]. A research on chronic obstructive pulmonary disease (COPD) developed a multilevel logistic model to generate small-area estimates of the prevalence of COPD in different geographic unites [47]. An analysis of the drinking pattern used a spatiotemporal model to estimate county-level alcohol use prevalence in the U.S [48]. This method considered spatial and temporal information as covariates to improve the predictions of all areas, as long as the area with limited sample sizes.

Censored count data were not analyzed particularly until 1985 when the censored Poisson regression model was developed and proposed by using the Newton-Raphson

algorithm to estimate unknown parameters [49]. Famoye and Wang (2004) expanded the censored Poisson regression model to the censored generalized Poisson regression model to handle censored data with over-dispersion or under-dispersion [50]. They used an iterative algorithm to get the maximum likelihood estimators, but they did not specify the iterative algorithm used. Mahmoud (2010) also developed a censored generalized Poisson regression model using a method similar to Famoye and Wang's method by adopting the Newton-Raphson algorithm [51].

However, as the log likelihood function of the censored generalized Poisson regression model has no close form to obtain the maximum likelihood estimates of unknown parameters, the Newton-Raphson algorithm cannot guarantee estimates reaching convergence because of the nonlinear nature of a problem [52]. Along with the application of Bayesian inference, Markov chain Monte Carlo (MCMC) simulation becomes a surrogate of the Newton-Raphson algorithm when priors can be pre-determined [53]. MCMC is computationally intensive in complex models, an approximate Bayesian inference called the integrated nested Laplace approximation (INLA), was developed to provide a more efficient algorithm with a lower computational burden [53]. The INLA method was originally used in latent Gaussian models. Along with advanced developments, this method has been extended to deal with other statistical models with a faster estimating speed and less biases [54]. Fong et al. (2010) performed INLA to estimate the parameters in generalized linear mixed models, and compared the estimation results with the ones using penalized quasi-likelihood in longitudinal datasets in the number of seizures[55]. Martins and Rue (2012) extended INLA to fit the spatial and spatial-temporal models in which the independent variable is no longer

Gaussian distributions [56]. Blangiardo et al. (2013) performed INLA to spatial and spatial-temporal models to analyze the suicides in London [57].

Public Health Significance

In public health, count datasets are commonly used to report statistics related to emerging or existing health problems, such as the prevalence and trends of various diseases, e.g. Lyme disease, Sexually Transmitted disease. As censored count data commonly occur in county-level health outcomes and in public health survey data, we need to find the proper methods to deal with the censored count data. As the complement of the data collecting, county-level datasets, or even smaller geographical area datasets (e.g., census group-level datasets), are always very large. Therefore, computational efficiency also needs to be improved. In the current research project, we compared three methods for dealing with censored count data under different censored proportions to determine which method has the best accuracy. In addition, we compared the computational time for each method to find the most efficient method to deal with censored count data. Therefore, the most accurate, time-efficient method helps us to prevent diseases by better determining the risk factors and identifying the patterns of disease.

Specific Aims

The major objectives of this study were to find and develop some proper methods to deal with the censored count data, then compare the accuracy and efficiency of each method under different censored proportions. Three specific aims of this dissertation were:

Aim 1:

To compare the results of censored Poisson regression model with the results of complete data using Poisson regression model, we need to make the censored part of count data complete. To check the performance of Poisson regression with imputed censored part, we compared the results of different censored proportion. Multiple imputation method was applied to the censored count data. After imputation the censored part, Poisson regression model was manipulate to do the parameter estimation.

Aim 2:

To improve the accuracy of the results from Poisson regression model with censored data, some geographical information related to the censored data was considered. Therefore, small area estimation method was applied to impute the censored count data with spatial properties.

Aim 3:

To improve the performance of censored Poisson regression in comparison with Poisson regression with complete data, new algorithms to estimate the unknown parameters for the censored Poisson regression model were developed. To avoid the convergence problem caused by Newton-Raphson algorithm, INLA method was developed. We expected more accurate results and more computational efficiency in censored Poisson regression model, especially with INLA algorithm.

METHODS

Methods for Specific Aim 1

The multiple imputation method uses a set of values replacing the missing values instead of using a single value for each missing datum. The “mi” or “mice” R packages are usually used to perform multiple imputation. In these R packages, Bayesian models were used to impute the data more precisely by giving multiple values than single values. Based on different properties for different data type, R packages provided different functions, i.e., if the data were binary, *mi.binary()* function was used; if the data were count data, *mi.count()* function was used.

We used multiple imputation to impute the censored count instead of missing data. We applied multiple imputation to simulated censored count dataset and real-world count dataset. We performed the following steps to impute the censored count data:

Step 1: Simulated the censored values independently using the estimated mean vector and covariance matrix. Censored observation (observation without values) was represented by $Y_{i(cen)}$, While the observation with value was represented by $Y_{i(obs)}$. An imputed $Y_{i(cen)}$ was drawn from a conditional distribution $Y_{i(cen)}|Y_{i(obs)}$.

Step 2: Simulated the posterior population mean vector and covariance matrix X from the complete sample estimates using a non-informative prior, which was built in R packages.

Step 3: Repeated step 1 and step 2 for 5 times as recommended by Rubin (1987) [58], which was built in R packages.

Step 4: Used the Poisson regression model with covariance matrix X and Y to impute the censored observation $Y_{i(cen)}$.

Step 5: Averaged the values and the standard errors of the parameter estimations across the censored value samples in order to obtain a single point estimate.

R packages “*mi*” and “*mice*” were applied to generate the values for the censored outcomes in order to produce the complete datasets. After generating the complete datasets, we used the Poisson regression model to estimate the coefficients of covariate matrix X.

Methods for specific Aim 2

The small area estimation method is a method commonly used in spatial analysis to estimate the unknown value for small counties. The idea of small area estimation uses the known observations of an outcome to estimate the unknown observations by the stratified demographic variables (e.g., age, gender, and race). We used small area estimation to estimate the censored count data by stratified demographic variables.

We assumed Y_{ijkc} as the count of an outcome variable (e.g. the number of cases) at age group i , race j , gender k in county c , which follows a Poisson distribution with a mean of λ_{ijkc} . Thus, the model was specified as follows,

$$\log(\lambda_{ijkc}) = \alpha + \beta_{1i} + \beta_{2j} + \beta_{3k} + f_{spat}(c) + \log(n_{ijkc}), \quad (1)$$

where β_{1i} , β_{2j} , and β_{3k} were fix effects for age, race, and gender, respectively. The spatial function $f_{spat}(c)$ was Markov random fields following an intrinsic conditional

autoregressive prior [61]. The last term $\log(n_{ijkc})$ was an offset corresponding to the logarithm of the at-risk population index by i, j, k , and c (the total number of individuals corresponding to Y_{ijkc}). We applied INLA algorithm described in Cadwell et al. [45] to estimate the censored count data.

We defined N_{ijkc} and Y_{ijkc} in equation (1) as age-race-gender-county specific at-risk population and those with the specific outcome, respectively. Thus, we derived Z_{ijkc} , the number of unobserved individuals (the censored cases) with the specific outcome, indexed by age, race, gender, and county straightforwardly. The sum of the observed and unobserved cases, $Y_{ijkc} + Z_{ijkc}$, was the total count of the outcome, where

$$Z_{ijkc} | Y_{ijkc}, n_{ijkc}, N_{ijkc} \sim \text{Poisson}(\mu_{ijkc}).$$

The parameter Y_{ijkc} was defined as

$$Y_{ijkc} = \left(\frac{\hat{\lambda}_{ijkc}}{n_{ijkc}} \right) \times (N_{ijkc} - n_{ijkc}) = \frac{\exp(\hat{\alpha} + \hat{\beta}_{1i} + \hat{\beta}_{2j} + \hat{\beta}_{3k} + \hat{f}_{spat}(c))}{n_{ijkc}} \times (N_{ijkc} - n_{ijkc}).$$

We applied the small area estimation method to estimate the censored count data, and then analyzed the data with the Poisson regression model to estimate the coefficient of all the covariates.

Method for Specific Aim 3

Multiple imputation method and small area estimation both estimate the censored count data to generate the complete dataset, but censored Poisson regression model can deal with censored count data directly without estimating the censored part.

We derived the censored Poisson regression model from the Poisson regression model. Suppose a response variable Y_i represents count data, and follows a Poisson distribution with a parameter λ_i , which can be predicted by covariates $\mathbf{x}_i = (x_1, x_2, x_3 \dots)$. For the Poisson regression model, we have the following equations:

$$\log \lambda = \boldsymbol{\beta} \mathbf{x} + \log E$$

$$y_i | \alpha, \boldsymbol{\beta}, \mathbf{x}_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = E \cdot e^{\alpha + \boldsymbol{\beta} \mathbf{x}_i},$$

where E is a constant of offset. The parameter λ_i equals the expectation and variance of Poisson distribution:

$$E(Y_i | \mathbf{x}) = \lambda_i$$

$$\text{Var}(Y_i | \mathbf{x}) = \lambda_i.$$

Censoring occurs when the value of Y_i is less than a constant C , so we define an indicator variable z_i as

$$z_i = \begin{cases} 1 & \text{if } Y_i \leq C \\ 0 & \text{otherwise} \end{cases},$$

and the probability of $z_i = 1$ is:

$$P(z_i = 1) = P(Y_i \leq C) = \sum_{y_i=0}^c p(\lambda_i) = \sum_{y_i=0}^c \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = Q(y_i)$$

$$f(y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

The likelihood function of left-censored Poisson regression model is

$$f(\mathbf{y}|\alpha, \beta) = \prod_{i=1}^n \left(\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right)^{1-z_i} (Q(y_i))^{z_i}, \quad (2)$$

and the log-likelihood function is

$$\log f(\mathbf{y}|\alpha, \beta) = \sum_{i=1}^n \left\{ (1 - z_i) [-\lambda_i + y_i \log \lambda_i - \log(y_i!)] + z_i [-\lambda_i + \log \left(\sum_{y_i=0}^c \frac{\lambda_i^{y_i}}{y_i!} \right)] \right\}.$$

We applied INLA algorithm to estimate the unknown parameter in equation (2).

MCMC algorithm also could be used to estimate the parameter, however, it encountered slow convergence and numerical instabilities. The procedure of MCMC was shown in Appendix A.

We first defined the posterior distribution of β given \mathbf{y} as

$$p(\beta|\mathbf{y}) = \frac{p(\alpha, \beta|\mathbf{y})}{p(\alpha|\beta, \mathbf{y})} \propto \frac{f(\mathbf{y}|\alpha, \beta)p(\alpha)p(\beta)}{p(\alpha|\beta, \mathbf{y})} \quad (3)$$

, and the priors of (α, β) were given by two normal distributions $N(\mu_{\alpha 0}, \sigma_{\alpha 0}^2)$ and

$N(\mu_{\beta 0}, \sigma_{\beta 0}^2)$, respectively. According to the Laplace approximation properties, the terms

depending on β in the numerator and denominator (Eq. 3) can cancel out. Thus, we fixed and

chose any arbitrary value for α in Eq. (3) and a convenient choice is $\alpha = \alpha_n$ as describe in Blangiardo's paper [62]. In order to evaluate the posterior distribution, some β are chosen based on the grid strategy included in the set $\{\beta^{(j)}\}$ and the value of density function is computed for each of them as the methods describe in Blangiardo's et al. [62],

$$p(\beta^{(j)}|y) \propto p(y|\alpha = \alpha_n, \beta = \beta^{(j)}) \cdot p(\alpha = \alpha_n) \cdot p(\beta^{(j)})$$

Then we evaluated of the full conditional distribution $p(\alpha|\beta, \mathbf{y})$ for each value of β in $\{\beta^{(j)}\}$ and of α in the set of $\{\alpha^{(l)}\}$. Thus, we evaluated $p(\alpha = \alpha^{(l)}|\beta = \beta^{(j)}, \mathbf{y})$. We estimated the marginal posterior distribution $p(\alpha|\mathbf{y})$ by integrating out β from the joint posterior $p(\alpha, \beta|\mathbf{y})$ through a finite weighted mean as the following equation, where $\Delta_j =$

$$\frac{1}{\sum_j p(\beta^{(j)}|y)}.$$

$$p(\alpha = \alpha^{(l)}|y) \propto \sum_j p(\alpha = \alpha^{(l)}|\beta = \beta^{(j)}, \mathbf{y}) p(\beta = \beta^{(j)}|y) \Delta_j.$$

We used the R package “INLA” to do the parameter estimations.

Simulation Study

Simulation Procedure

First, we simulated the censored count dataset for all three methods using the following steps.

Step 1: Created the independent variables (covariate matrix X) in the censored count dataset. We assumed that there were 3 independent variables in the covariate matrix X in our

data, which were X_1 , X_2 , and X_3 . Then we assumed that X_1 , X_2 , and X_3 followed normal distributions, which were $N(0.05, 0.2)$, $N(0.1, 0.1)$, and $N(0.2, 0.05)$, respectively. We simulated 1000 values in each X_1 , X_2 , and X_3 , so there were 1000 rows and 3 columns (X_1 , X_2 , and X_3) in our covariate matrix X .

Step 2: Created the outcome variable (Y) in the censored count dataset. The outcome variable Y followed *Poisson* (λ) distribution and $\lambda = \exp(b_1 + b_2x_{1i} + b_3x_{2i} + b_4x_{3i} + offset) = \exp(bX + offset)$. Offset was the population of the sample size, and in our simulation the offset was a constant. When we set up the true value b_1, b_2, b_3 , and b_4 as 1, 2, 3, and 4, respectively, we generated our outcome count variable Y based on true value of b and covariate matrix X . We created 1000 values of Y corresponding to 1000 rows of covariate matrix X . By doing this, we obtained the complete count dataset including Y , X_1 , X_2 , and X_3 , and each variable had 1000 rows. Therefore, the complete dataset was a matrix with 1000 rows and 4 columns (Y , X_1 , X_2 , and X_3).

Step 3: Created censored count dataset. We chose three censored points, which were 7, 10, and 12, in order to make the censored proportion to be around 10%, around 30%, and around 50%, respectively. When Y_i was less than the censored point c , which means Y_i is censored, we made Y_i to be the smallest value in censored region ($Y_i = 1$). Then we assigned an indicator Z : $Z=1$ represented Y_i censored ($Y_i < c$), otherwise, $Z=0$. Then we saved the censored count datasets, including censored count outcome Y , independent variables X_1 , X_2 , X_3 and indicator variable Z . Therefore, the censored count dataset was a matrix of 1000 rows and 5 columns (Y , X_1 , X_2 , X_3 , and Z).

Step 4: Repeated Step 1 to Step 3 for 1000 times to create 1000 censored datasets.

Second, after creating the censored count datasets, we applied multiple imputation method, small area estimation method and censored Poisson regression model method to deal with censored count dataset. Then we saved the estimations of each parameter and we compared the accuracy of three methods, by using the relative error and bar plots. Relative error was represented as follows

$$Relative\ error = \frac{|estimates - true\ value|}{true\ value} * 100\%,$$

and bar plots were used to compare the coefficients.

Application of Multiple Imputation to Simulated Censored Count Datasets

We applied R packages “mi” and “mice” to perform the multiple imputation on the simulated censored count datasets. Function *mice()* was applied to impute data, then *pool()* function was used to take average of imputed data. After generating the complete dataset, we applied Poisson regression model to the complete data and estimated the parameters for covariate matrix, whose true values were b_1, b_2, b_3 , and b_4 . We calculated the relative error for each estimation, then we saved the estimation results for accuracy comparison.

Application of Small Area Estimation to Simulated Censored Count Datasets

We applied small area estimation, which used the existing individual demographic variables to estimate the unknown outcome observations. We completed the following steps:

Step 1: Used the demographic variables to estimate the censored count data. We used stratified demographic variables, such as age group and race, to build up the model to

estimate the censored count data. We used R packages “*inla*” to do the small area estimation. We used function *inla()* to estimate the full dataset of outcome dataset, then we extracted the fitted value from demographic model. According to the indicator variable *Z*, we identified the censored outcome data.

Step 2: Completed censored outcome data. We tried the following two ways to complete the censored part of the dataset. First, we used the estimated *Y*, even if it exceeded the censored point. For example, when the censored point was 7, the values lower than 7 were censored. When the fitted outcome was 8, we still used 8 to impute the censored value, even though 8 was larger than 7. Second, when the estimated *Y* exceeded the censored point, we used the censored point instead of the estimated *Y*. For example, when the censored point was 7, but the imputed value was 8 (larger than 7), we used 7 (the censored point) to impute the values instead of 8 (the fitted value).

Step 3: Estimated the coefficients for covariate matrix. We used Poisson regression model to estimate the coefficient of independent variables. We used *glm()* function to manipulate Poisson regression model and estimated the parameters, whose true values were b_1 , b_2 , b_3 , and b_4 . We repeated 1000 datasets to calculate the average of mean and standard deviation for each coefficient.

We plotted bar plots to show the value of different estimation for small area estimation under different censored proportions, which were 7, 10, and 12 mentioned in Section 2.4.1. We compared the accuracy of results of parameter estimations by calculating the relative errors.

Application of Censored Poisson Regression to Censored Count Dataset

We used R package “INLA” and directly applied a censored Poisson regression model to the existing censored count datasets. We used INLA algorithm to estimates the coefficients (b_1 , b_2 , b_3 , and b_4), whose true values were 1, 2, 3, and 4, respectively. Then we compared the accuracy of this method using relative error, and plotted bar charts to depict the estimation results.

Case study

Data Description

We wanted to compare the performance of multiple imputation, small area estimation, and censored Poisson regression dealing with censored count data using real-world data. Thus, we applied these three methods to assess the association between heat wave and cardiovascular diseases using hospital admission data from Harris County, Texas, from 2006 to 2011. In this dataset, we collected the individual-level data for every admission. We collected the county-level complete count dataset for cardiovascular disease. In this analysis, we considered the estimations using the complete count dataset as our “true value”, so that we were able to compare the estimation results for censored count data with the true value (estimations under the complete results).

First, we collected hospital admissions data for the period 2006-2011 from the Texas Department of State Health Services. Hospital admission data are individual-level data, including gender, age, race, record ID, diagnosis code, patients’ home address and type of admission. The diagnosis code represented different diseases (health outcomes) that caused hospital admission, and were based on the International Statistical Classification of Diseases and Related Health Problems 9th Revision code (known as ICD-9 code). Thus, different diseases had different ICD-9 codes, ranging from 100.000 to 999.999. ICD-9 code for cardiovascular diseases ranged from 390 to 429. Therefore, if the diagnostic code (ICD-9 code) fell between 390 and 429, we defined the variable “Cardiovascular” as 1 for the corresponding individual, otherwise, “Cardiovascular” was defined as 0.

Second, we aggregated hospital admission data for cardiovascular outcomes by Federal Information Processing Standard (FIPS) codes at the county level and by admission date based on the patient who was diagnosed with cardiovascular disease. After aggregating, there were a total of 2,191 rows in Harris County data. The number of individuals who were diagnosed with cardiovascular disease by county by date was the outcome in our model. We defined different censored values, so that the data had different censored proportion based on different censored values. We compared the results under different censored proportions with the results using the complete dataset in order to compare the accuracy of different methods dealing with censored count data.

Third, we obtained weather data from the National Center for Climatic Center through the Integrated Surface Database [60]. We used the maximum temperature for each day in Harris County to analyze how the highest temperature for each day affected the cardiovascular outcomes. There were no missing values in maximum temperature data.

Lastly, we combined the daily hospital admission data with weather extreme data by FIPS code and date to get the complete count dataset. In the complete count dataset, our health outcome of interest was the number of individuals who were diagnosed with cardiovascular disease was the outcome of interest, while the maximum temperature was the covariate in the dataset. Because heat waves may affect diseases with several days of delay, we added three lag terms in our model. For example, when the lag term equals to 1, the heat wave affects the disease one day after the heat wave. In addition, time should be considered in our data, and since time could not be taken as linear term, we added polynomial term for time effect. In sum, in our case study, the outcome was the number of individuals who were

diagnosed with cardiovascular disease for Harris County by date, and the covariates were the maximum temperature, the three lag terms for heat effect, and three polynomial terms for time.

Data Analysis

Before data analysis, we chose different censored points to create different censored proportions, so we could compare the estimation results under different censored proportions. In this analysis, we chose the censored points equal to 7, 8, 9, 10, 11, 12, 13, and 14, then we used these 8 datasets to calculate the estimates and compared the results.

We applied multiple imputation and small area estimation to impute the censored part to 8 datasets with different censored proportions. Then we used Poisson regression model to estimate the coefficients of 7 covariates, which are maximum heat, lag 1 for heat, lag 2 for heat, lag 3 for heat, polynomial term time 1, polynomial term time 2, and polynomial term time 3. For the censored Poisson regression model, we directly applied the INLA algorithm to estimate the coefficients of the 7 covariates.

To justify the results, we also estimated the coefficients of 7 covariates under the condition of complete datasets, which were taken as the true values during comparison. We applied Poisson regression model to complete count data to estimate the coefficients under the perfect condition (no censored, no missing values in the count datasets). Also, we used the “old method” of dealing with censored count data, which considers the censored values as missing values. In this method, we directly applied the Poisson regression model to censored count data to estimate the coefficients.

We plotted the bar plot for each estimated coefficient under 8 different censored proportion of following 5 models: (1) Poisson regression model under complete datasets; (2) Poisson regression model directly applied to censored count data (censored data with Poisson regression model, short as CDPRM); (3) multiple imputation method dealing with censored count data (MI method); (4) small area estimation method dealing with censored count data (SAE); (5) censored Poisson regression model (CPRM).

Limitations

This study has some important limitations. First, we did not identify trends or pattern in the fluctuation of relative errors as the censored proportion increased across all methods. Second, we used a real-world censored count datasets may generate results than other real-world censored count datasets. Thus, we only provided a reference instead of a standard for determining the best performance under different specific censored proportions. In the future, these methods could be applied to additional real-world censored count datasets to identify a relatively precise standard for different censored proportions. Third, we used censored count datasets that met Poisson distribution assumption, which provided better simulation results than real-world results. In the future, more accurate models to deal with censored count data with less distribution assumptions could be developed.

Strengths

Despite the limitations, this study has some strengths. First, to our knowledge, this study is the first to compare all the methods dealing with censored count data considering both estimation accuracy and computational efficiency simultaneously. Thus, we identified

the best method under different censored proportions, which provides a means for determining the best method to use based on the size of the censored count dataset and of the censored proportion. Second, we improved computational efficiency of CPRM by using INLA algorithm rather than the other typically used algorithms (i.e., MCMC and Newton Raphson) [61] for parameter estimates.

APPENDICES

Appendix A

MCMC Method to estimate the parameters in censored Poisson regression model

Among Monte Carlo Markov Chain (MCMC) methods, we adopted the Metropolis-Hasting (M-H) algorithm to estimate the unknown parameters (α, β) . In M-H algorithm, equation (3) was the target density f , equation (4) was used during the calculation. The candidate densities q are independent normally distributed, which were the priors of (α, β) .

M-H algorithm was as follow:

1. Generated initial values (α^0, β^0) from two normal distributions, $N(\hat{\alpha}, var(\hat{\alpha}))$ and $N(\hat{\beta}, var(\hat{\beta}))$, which could be gained from the regular Poisson regression model.
2. Accepted this candidate with a probability

$$r(\alpha^0, \beta^0) = \min\left(\frac{L(\alpha^{(0)}, \beta^{(0)}|\mathbf{y})q(\alpha^{(t-1)}, \beta^{(t-1)})}{L(\alpha^{(t-1)}, \beta^{(t-1)}|\mathbf{y})q(\alpha^{(0)}, \beta^{(0)})}, 1\right).$$

3. Generated a random number u from a uniform distribution $U(0, 1)$.
4. If $u \leq r(\alpha^0, \beta^0)$, set $(\alpha^{(t)}, \beta^{(t)}) = (\alpha^{(0)}, \beta^{(0)})$. Otherwise, set $(\alpha^{(t)}, \beta^{(t)}) = (\alpha^{(t-1)}, \beta^{(t-1)})$

, where t represented the times of iterations, $(\alpha^{(t)}, \beta^{(t)})$ mean the estimated (α, β) from the t th iteration. L was the likelihood function, which had the same function as f . r was the probability of accepting the candidates. u was a criterion for judging if we could stop the circulation.

REFERENCES

- [1] Cordero A, Mulinare J, Berry R, Boyle C, Dietz W, Johnston Jr R, et al. CDC Grand Rounds: additional opportunities to prevent neural tube defects with folic acid fortification. *Morbidity and mortality weekly report*. 2010;59:980-4.
- [2] Prevention CfDCa. Lyme Disease Home. 2017.
- [3] Frome E. PREG: A computer program for Poisson regression analysis. Oak Ridge Associated Universities, Inc., TN (USA); 1981.
- [4] Bosch FX, Manos MM, Muñoz N, Sherman M, Jansen AM, Peto J, et al. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. *Journal of the National cancer institute*. 1995;87:796-802.
- [5] Frome EL. The analysis of rates using Poisson regression models. *Biometrics*. 1983:665-74.
- [6] Cox S, West SG, Aiken LS. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of personality assessment*. 2009;91:121-36.
- [7] Frome E, Watkins J, Ellis E. Poisson regression analysis of illness and injury surveillance data. Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN (United States); 2012.
- [8] Gupta PL, Gupta RC, Tripathi RC. Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis*. 1996;23:207-18.
- [9] Yau KK, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in medicine*. 2001;20:2907-20.
- [10] Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Zero-truncated and zero-inflated models for count data. *Mixed effects models and extensions in ecology with R*: Springer; 2009. p. 261-93.
- [11] Nakaya T, Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted Poisson regression for disease association mapping. *Statistics in medicine*. 2005;24:2695-717.
- [12] Ridout MS, Besbeas P. An empirical model for underdispersed count data. *Statistical Modelling*. 2004;4:77-89.

- [13] Best NG, Ickstadt K, Wolpert RL. Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American statistical association*. 2000;95:1076-88.
- [14] Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. *Journal of statistical software*. 2008;27:1-25.
- [15] Rosamond W, Flegal K, Furie K, Go A, Greenlund K, Haase N, et al. Heart disease and stroke statistics—2008 update. *Circulation*. 2008;117:e25-e146.
- [16] Sellers KF, Shmueli G. Predicting censored count data with COM-Poisson regression. 2010.
- [17] McShane B, Adrian M, Bradlow ET, Fader PS. Count models based on Weibull interarrival times. *Journal of Business & Economic Statistics*. 2008;26:369-78.
- [18] data.illinois.gov. IDPH STD Illinois By County Rank. 2017.
- [19] Sellers KF, Borle S, Shmueli G. The COM-Poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*. 2012;28:104-16.
- [20] van der Heijden GJ, Donders ART, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology*. 2006;59:1102-9.
- [21] Whitehead NS, Leiker R. Case management protocol and declining blood lead concentrations among children. *Prev Chronic Dis* [serial online]. 2007.
- [22] Dignam TA, Lojo J, Meyer PA. Reduction of elevated blood lead levels in children in North Carolina and Vermont, 1996–1999. *Environmental health perspectives*. 2008;116:981.
- [23] Schmee J, Hahn GJ. A simple method for regression analysis with censored data. *Technometrics*. 1979;21:417-32.
- [24] Shih WJ. Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine*. 2002;3:4.
- [25] Jin X, Carlin BP, Banerjee S. Generalized hierarchical multivariate CAR models for areal data. *Biometrics*. 2005;61:950-61.
- [26] Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581-92.
- [27] Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004.

- [28] Rubin DB, Schenker N. Multiple imputation in health-care databases: An overview and some applications. *Statistics in medicine*. 1991;10:585-98.
- [29] Taylor JM, Muñoz A, Bass SM, Saah AJ, Chmiel JS, Kingsley LA. Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation. *Statistics in Medicine*. 1990;9:505-14.
- [30] Dorey F, Little R, SCHENKER N. Multiple imputation for interval-censored threshold data. *Joint Statistical Meetings, Anaheim*1990.
- [31] Royston P. Multiple imputation of missing values. *Stata journal*. 2004;4:227-41.
- [32] Royston P. Multiple imputation of missing values: update of ice. *Stata Journal*. 2005;5:527.
- [33] Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. *Stata Journal*. 2009;9:466.
- [34] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*. 2011;30:377-99.
- [35] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*. 2009;338:b2393.
- [36] Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. *Statistics in medicine*. 2001;20:1541-9.
- [37] Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*. 1999;18:681-94.
- [38] Schenker N, Raghunathan TE, Chiu P-L, Makuc DM, Zhang G, Cohen AJ. Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*. 2006;101:924-33.
- [39] Bartlett JW, Seaman SR, White IR, Carpenter JR, Initiative* AsDN. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research*. 2015;24:462-87.
- [40] AÄŸmann C, WÄ¼rbach A, GoÄŸmann S, Geissler F, Bela A. Nonparametric Multiple Imputation for Questionnaires with Individual Skip Patterns and Constraints: The

Case of Income Imputation in the National Educational Panel Study. *Sociological Methods & Research*. 2017;46:864-97.

[41] Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data. *Jama*. 2015;314:1966-7.

[42] Eekhout I, de Vet HC, Twisk JW, Brand JP, de Boer MR, Heymans MW. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of clinical epidemiology*. 2014;67:335-42.

[43] Enders CK, Mistler SA, Keller BT. Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological methods*. 2016;21:222.

[44] Ghosh M, Rao J. Small area estimation: an appraisal. *Statistical science*. 1994;55-76.

[45] Cadwell BL, Thompson TJ, Boyle JP, Barker LE. Bayesian small area estimates of diabetes prevalence by US county, 2005. *Journal of Data Science*. 2010;8:171-88.

[46] Barker LE, Thompson TJ, Kirtland KA, Boyle JP, Geiss LS, McCauley MM, et al. Bayesian small area estimates of diabetes incidence by United States county, 2009. *Journal of data science: JDS*. 2013;11:269.

[47] Zhang X, Holt JB, Lu H, Wheaton AG, Ford ES, Greenlund KJ, et al. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American journal of epidemiology*. 2014;179:1025-33.

[48] Dwyer-Lindgren L, Flaxman AD, Ng M, Hansen GM, Murray CJ, Mokdad AH. Drinking patterns in US counties from 2002 to 2012. *American journal of public health*. 2015;105:1120-7.

[49] Terza JV. A Tobit-type estimator for the censored Poisson regression model. *Economics Letters*. 1985;18:361-5.

[50] Famoye F, Wang W. Censored generalized Poisson regression model. *Computational statistics & data analysis*. 2004;46:547-60.

[51] Mahmoud MM, Alderiny MM. On estimating parameters of censored generalized Poisson regression model. *Applied Mathematical Sciences*. 2010;4:623-35.

- [52] Veleba J. Stability Algorithms for Newton-Raphson Method in Load Flow Analysis. Intensive Programme “Renewable Energy Sources. 2010.
- [53] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*. 2009;71:319-92.
- [54] Martins TG, Simpson D, Lindgren F, Rue H. Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*. 2013;67:68-83.
- [55] Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. *Biostatistics*. 2010;11:397-412.
- [56] Martins TG, Rue H. Extending INLA to a class of near-Gaussian latent models. *arXiv preprint arXiv:12101434*. 2012.
- [57] Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*. 2013;7:39-55.
- [58] Rubin DB. Comment. *Journal of the American Statistical Association*. 1987;82:543-6.
- [59] Quality TCoE. Texas Air Monitoring Information System. 2014.
- [60] information NCfe. Integrated Surface Database. 2016.
- [61] Kindermann RP, Snell JL. On the relation between Markov random fields and social networks. *Journal of Mathematical Sociology*. 1980;7:1-13.
- [62] Blangiardo M, Cameletti M. Spatial and spatio-temporal Bayesian models with R-INLA: John Wiley & Sons; 2015.

JOURNAL ARTICLE #1

Development of Integrated Laplace Approximation in the Censored Poisson Regression Model

Journal of Statistical Computation and Simulation

Yu X¹, Chan W¹, Swint JM¹, Zhang K¹, Chien LC².

¹ *School of Public Health, University of Texas Health Science Center at Houston, TX, USA*

² *Community Health Sciences, University of Nevada, Las Vegas, NV, USA*

Abstract

Background: Poisson regression models are commonly used in the statistical analysis of count data, but sometimes data are censored. Previous work usually defined the censored part as the missing data and removed them from statistical analysis, which reduced the power apparently.

Methods: The censored Poisson model was developed to analyze the censored count data, but the estimating algorithm is hard to reach convergence in practical research. This study developed algorithms based on integrated nested Laplace approximation (INLA) for estimating unknown parameters in the censored Poisson model. We simulated the censored count datasets under different censored proportions and compared the results under different censored proportions with the results obtained from Poisson regression model using complete count dataset. We applied censored Poisson regression model and INLA algorithm to assess the association between heat wave and cardiovascular diseases using hospital admission data from Harris County, Texas, from 2006 to 2011

Results: We found that with the censored proportion getting larger, the estimated coefficients of INLA were still close to the true values. The average relative error for the estimation is under 1% even the censored proportion was greater than 50%.

Conclusions The INLA provides an efficient algorithm to conduct accurate estimates in the censored Poisson model regardless the proportions of censorship.

Key words: censored count data, censored Poisson regression model, INLA

Introduction

Poisson regression models are commonly used in the statistical analysis of count data, but sometimes data are censored when the number of count observations is below a certain value because of confidentiality. Different from data masking containing specific techniques to substitute, shuffle, or encrypt accessible data, censored data are hidden when they satisfy some conditions. For example, the Illinois Department of Public Health Sexually Transmitted Diseases dataset does not publish sexually transmitted disease data for counties with a population less than 15,000 or with a total birth rate less than 300, so the null values in the dataset are censored data [1]. Censored counts often appear in the CDC county-level datasets. In the Centers for Disease Control and Prevention Lead Poisoning Prevention Program surveillance database, the number of children with elevated blood lead levels in each county was censored by 5 or less [2].

Logically, censored data are not missing, but just intentionally masked when people want to access them. Hence, regarding censored data as missing data in statistical analysis may lead biased results. A study analyzed censored data defined as missing data, and concluded that the bias of estimations increases as the proportion of censored data increases [3]. Medical research in blood lead level reduction shows that the results of including the censored data are more reasonable than the ones excluding the censored data [4, 5].

Censorship has been taken care in survival analysis, whatever times to death [6] or times to infection [7], while the development in the Poisson regression model is still limited. In 1985, Terza first defined the censored Poisson regression model, and used the Newton-Raphson algorithm to estimate unknown parameters [8]. The author compared the results

with Frome's paper [6], in which excluded the censored data, and found a bias in excess of 100% when censoring ignored. A couple of studies had applied the Newton-Raphson method to estimate unknown parameters of censored generalized Poisson regression model [7].

Because the log likelihood function of the censored Poisson model has no close form to obtain the maximum likelihood estimates of unknown parameters, applying an iteration algorithm is reasonable. The Newton-Raphson method can conveniently find better approximations to the zero of a log likelihood function; however, estimates may not reach convergence because of the nonlinear nature of a problem [9]. Along with the application of Bayesian inference, Monte Carlo Markov Chain (MCMC) simulation technique becomes a surrogate of the Newton-Raphson method when priors can be pre-determined [10]. However, MCMC is computationally intensive in complex models; thus, a new method called integrated nested Laplace approximation (INLA) was developed to provide a more efficient algorithm with a lower computational burden [10].

In this paper, we developed the censored Poisson regression model to deal with censored count data under different censored proportions, and used INLA methods to estimate unknown parameters. A simulation was proposed to compare the performance of INLA algorithm with true values under different censored proportions. In addition, we applied INLA algorithm in censored Poisson regression model to real-world censored count datasets under different censored proportions to assess the association between heat wave and cardiovascular diseases using hospital admission data from Harris County, Texas, from 2006 to 2011.

Methods

Statistical Model and Parameter Estimation

We derived the censored Poisson regression model from the Poisson regression model. Suppose a response variable Y_i represents count data, and follows a Poisson distribution with a parameter λ_i , which can be predicted by covariates $\mathbf{x}_i = (x_1, x_2, x_3 \dots)$. For the Poisson regression model, we have the following equations:

$$\log \lambda = \boldsymbol{\beta} \mathbf{x} + \log E$$

$$y_i | \alpha, \boldsymbol{\beta}, \mathbf{x}_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = E \cdot e^{\alpha + \boldsymbol{\beta} \mathbf{x}_i},$$

where E is a constant of offset. The parameter λ_i equals the expectation and variance of Poisson distribution:

$$E(Y_i | \mathbf{x}) = \lambda_i$$

$$\text{Var}(Y_i | \mathbf{x}) = \lambda_i.$$

Censoring occurs when the value of Y_i is less than a constant C , so we define an indicator variable z_i as

$$z_i = \begin{cases} 1 & \text{if } Y_i \leq C \\ 0 & \text{otherwise} \end{cases},$$

and the probability of $z_i = 1$ is:

$$P(z_i = 1) = P(Y_i \leq C) = \sum_{y_i=0}^c p(\lambda_i) = \sum_{y_i=0}^c \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = Q(y_i)$$

$$f(y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

The likelihood function of left-censored Poisson regression model is

$$f(\mathbf{y}|\alpha, \beta) = \prod_{i=1}^n \left(\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right)^{1-z_i} (Q(y_i))^{z_i}, \quad (2)$$

and the log-likelihood function is

$$\log f(\mathbf{y}|\alpha, \beta) = \sum_{i=1}^n \left\{ (1 - z_i) [-\lambda_i + y_i \log \lambda_i - \log(y_i!)] + z_i [-\lambda_i + \log \left(\sum_{y_i=0}^c \frac{\lambda_i^{y_i}}{y_i!} \right)] \right\} \quad (3).$$

We applied INLA algorithm to estimate the unknown parameter in equation (3). We first defined the posterior distribution of β given y as

$$p(\beta|y) = \frac{p(\alpha, \beta|\mathbf{y})}{p(\alpha|\beta, \mathbf{y})} \propto \frac{f(\mathbf{y}|\alpha, \beta)p(\alpha)p(\beta)}{p(\alpha|\beta, \mathbf{y})} \quad (3)$$

, and the priors of (α, β) were given by two normal distributions $N(\mu_{\alpha 0}, \sigma_{\alpha 0}^2)$ and

$N(\mu_{\beta 0}, \sigma_{\beta 0}^2)$, respectively. According to the Laplace approximation properties, the terms

depending on β in the numerator and denominator (Eq. 3) can cancel out. Thus, we fixed and

chose any arbitrary value for α in Eq. (3) and a convenient choice is $\alpha = \alpha_n$ as describe in

Blangiardo's paper [11]. In order to evaluate the posterior distribution, some β are chosen

based on the grid strategy included in the set $\{\beta^{(j)}\}$ and the value of density function is computed for each of them as the methods describe in Blangiardo's et al. [12],

$$p(\beta^{(j)}|y) \propto p(y|\alpha = \alpha_n, \beta = f(\beta^{(j)})) \cdot p(\alpha = \alpha_n) \cdot p(\beta^{(j)})$$

Then we evaluated of the full conditional distribution $p(\alpha|\beta, \mathbf{y})$ for each value of β in $\{\beta^{(j)}\}$ and of α in the set of $\{\alpha^{(l)}\}$. Thus, we evaluated $p(\alpha = \alpha^{(l)}|\beta = \beta^{(j)}, \mathbf{y})$. We estimated the marginal posterior distribution $p(\alpha|\mathbf{y})$ by integrating out β from the joint posterior $p(\alpha, \beta|\mathbf{y})$ through a finite weighted mean as the following equation, where $\Delta_j = \frac{1}{\sum_j p(\beta^{(j)}|y)}$:

$$p(\alpha = \alpha^{(l)}|y) \propto \sum_j p(\alpha = \alpha^{(l)}|\beta = \beta^{(j)}, \mathbf{y}) p(\beta = \beta^{(j)}|y) \Delta_j.$$

We used the R package “INLA” to do the parameter estimations.

Simulation Study

Figure 1 shows how we simulated the censored dataset in a flow chart. We simulated the censored count datasets using the following steps:

Step 1: Created the independent variables (covariate matrix X) in the censored count dataset. We assumed that there were 3 independent variables in the covariate matrix X in our data, which were X_1 , X_2 , and X_3 . Then we assumed that X_1 , X_2 , and X_3 followed normal distributions, which were $N(0.05, 0.2)$, $N(0.1, 0.1)$, and $N(0.2, 0.05)$, respectively. We simulated 1000 values in each X_1 , X_2 , and X_3 , so there were 1000 rows and 3 columns (X_1 , X_2 , and X_3) in our covariate matrix X .

Step 2: Created the outcome variable (Y) in the censored count dataset. The outcome variable Y followed *Poisson* (λ) distribution and $\lambda = \exp(b_1 + b_2x_{1i} + b_3x_{2i} + b_4x_{3i} + offset) = \exp(bX + offset)$. Offset was the population of the sample size, and in our simulation the offset was a constant. When we set up the true value b_1, b_2, b_3 , and b_4 as 1, 2, 3, and 4, respectively, we generated our outcome count variable Y based on true value of b and covariate matrix X . We created 1000 values of Y corresponding to 1000 rows of covariate matrix X . By doing this, we obtained the complete count dataset including Y, X_1, X_2 , and X_3 , and each variable had 1000 rows. Therefore, the complete dataset was a matrix with 1000 rows and 4 columns (Y, X_1, X_2 , and X_3).

Step 3: Created censored count dataset. We chose three censored points, which were 7, 10, and 12, in order to make the censored proportion to be around 10%, around 30%, and around 50%, respectively. When Y_i was less than the censored point c , which means Y_i is censored, we made Y_i to be the smallest value in censored region ($Y_i = 1$). Then we assigned an indicator Z : $Z=1$ represented Y_i censored ($Y_i < c$), otherwise, $Z=0$. Then we saved the censored count datasets, including censored count outcome Y , independent variables X_1, X_2, X_3 and indicator variable Z . Therefore, the censored count dataset was a matrix of 1000 rows and 5 columns (Y, X_1, X_2, X_3 , and Z).

Step 4: Repeated Step 1 to Step 3 for 1000 times to create 1000 censored datasets.

After creating 1000 censored count datasets, we applied censored Poisson regression to the 1000 simulated censored count datasets and used INLA algorithm to estimate the coefficient of covariate matrix X . Then we took the average of each estimations for 1000

datasets as the estimation of covariate matrix X and the average of standard errors of the parameter estimations as the standard deviation.

We used the relative error to compare the estimation results with true values for censored Poisson regression model. Relative error was represented as follows

$$Relative\ error = \frac{|estimates - true\ value|}{true\ value} * 100\%.$$

Case Study

We applied censored Poisson Regression model to assess the association between heat wave and cardiovascular diseases using hospital admission data from Harris County, Texas, from 2006 to 2011. In this dataset, we collected the individual-level data for every admission. We collected the county-level complete count dataset for cardiovascular disease. In this analysis, we considered the estimations using the complete count dataset as our “true value”, so that we were able to compare the estimation results for censored count data with the true value (estimations under the complete results).

First, we collected hospital admissions data for the period 2006-2011 from the Texas Department of State Health Services. Hospital admission data are individual-level data, including gender, age, race, record ID, diagnosis code, patients’ home address and type of admission. The diagnosis code represented different diseases (health outcomes) that caused hospital admission, and were based on the International Statistical Classification of Diseases and Related Health Problems 9th Revision code (known as ICD-9 code). Thus, different

diseases had different ICD-9 codes, ranging from 100.000 to 999.999. ICD-9 code for cardiovascular diseases ranged from 390 to 429. Therefore, if the diagnostic code (ICD-9 code) fell between 390 and 429, we defined the variable “Cardiovascular” as 1 for the corresponding individual, otherwise, “Cardiovascular” was defined as 0.

Second, we aggregated hospital admission data for cardiovascular outcomes by Federal Information Processing Standard (FIPS) codes at the county level and by admission date based on the patient who was diagnosed with cardiovascular disease. After aggregating, there were a total of 2,191 rows in Harris County data. The number of individuals who were diagnosed with cardiovascular disease by county by date was the outcome in our model. We defined different censored values, so that the data had different censored proportion based on different censored values. We compared the results under different censored proportions with the results using the complete dataset in order to compare the accuracy of different methods dealing with censored count data.

Third, we obtained weather data from the National Center for Climatic Center through the Integrated Surface Database [13]. We used the maximum temperature for each day in Harris County to analyze how the highest temperature for each day affected the cardiovascular outcomes. There were no missing values in maximum temperature data.

Lastly, we combined the daily hospital admission data with weather extreme data by FIPS code and date to get the complete count dataset. In the complete count dataset, our health outcome of interest was the number of individuals who were diagnosed with cardiovascular disease was the outcome of interest, while the maximum temperature was the

covariate in the dataset. Because heat waves may affect diseases with several days of delay, we added three lag terms in our model. For example, when the lag term equals to 1, the heat wave affects the disease one day after the heat wave. In addition, time should be considered in our data, and since time could not be taken as linear term, we added polynomial term for time effect. In sum, in our case study, the outcome was the number of individuals who were diagnosed with cardiovascular disease for Harris County by date, and the covariates were the maximum temperature, the three lag terms for heat effect, and three polynomial terms for time.

Before data analysis, we chose different censored points to create different censored proportions, so we could compare the estimation results under different censored proportions. In this analysis, we chose the censored points equal to 7, 8, 9, 10, 11, 12, 13, and 14, then we used these 8 datasets to calculate the estimates and compared the results.

We applied censored Poisson regression model to 8 datasets with different censored proportions. Then we used INLA algorithm to estimate the coefficients of 7 covariates, which are maximum heat, lag 1 for heat, lag 2 for heat, lag 3 for heat, polynomial term time 1, polynomial term time 2, and polynomial term time 3.

To justify the results, we also estimated the coefficients of 7 covariates under the condition of complete datasets, which were taken as the true values during comparison. We compared the estimation results from censored Poisson regression model with the estimation results from the complete datasets using relative errors, which were described in simulation study section.

Results

Simulation Study

Tables 1 presents the simulation results of censored Poisson regression method under different proportions of censored count datasets. We chose the censored point equal to 7, 10, and 12, resulting in censored proportion by 7.9%, 33.6%, and 54.1%, respectively.

When the censored proportion was 7.9%, 33.6%, and 54.1%, the average relative errors of parameter estimations were 0.21%, 0.32%, and 0.41%, respectively (shown in Table 1). When the censored proportion was increasing, the relative error of the parameter estimation was increasing. Although the censored proportion was greater than 50%, the relative error for censored Poisson regression model was still less than 1%. The standard deviation for each parameter estimation remained the same as the censored proportion was increasing.

Case Study

Table 2 – Table 3 present the case study results of the censored Poisson regression model to compare the accuracy of the parameter estimations under different censored proportions. We chose the censored points equal to 7, 8, 9, 10, 11, 12, 13, and 14, resulting in censored proportion by 4.56%, 8.08%, 12.96%, 20.41%, 28.07%, 37.88%, 48.65%, and 59.33%, respectively.

Table 1 presents the results when the censored proportion were 4.56%, 8.08%, 12.96%, and 20.41% comparing the results with complete data using censored Poisson regression model. The average relative errors of censored Poisson regression model were

20.99%, 21.07%, 16.21%, and 8.63%, respectively (shown in Table 3). The average relative error was around 16% when the censored proportion under 25% (data was not shown on the table). The standard deviation remained the same when the censored proportion was increasing. The signs of parameter estimation of time3 (polynomial term for time) were opposite from the true value when the censored proportions were greater than 12.96%. Since this polynomial term was an adjustment term for time effect and time 3 did not carry any concrete information, the opposite sign could be acceptable in this situation.

Table 3 presents the results when the censored proportion were 28.07%, 37.88%, 48.65%, and 59.33% comparing the results with complete data using censored Poisson regression model. The average relative errors of censored Poisson regression model were 78.91%, 154.38%, 151.21%, and 98.90%, respectively (shown in Table 3). Comparing with the results under the censored proportion less than 20.41%, the relative errors were dramatically increasing and fluctuated. The standard deviation still remained the same when the censored proportion was increasing. The signs of parameter estimation of time3 (polynomial term for time) were still opposite from the true value when the censored proportions were greater than 12.96%. The sign of Max Heat turned to the opposite side since the censored proportion was larger than 37.88%, which would provide an opposite effect when analyzing the association between maximum heat and hospitalization. Thus, the estimation results under the censored proportion greater than 37.88% were not reasonable in this real-world censored count dataset.

Discussion

In our simulation study, the simulated censored count data represented an ideal scenario, in which the outcome variable followed a Poisson distribution with a controlled mean and variance (the mean was equal to the variance). Under this condition, the censored Poisson regression model performed exceptionally well (relative error less than 1%) even when the censored proportion was greater than 50% censored proportion.

In our case study, we assessed the association between heat wave temperature and hospitalization due to cardiovascular diseases in Harris County, Texas, from 2006 to 2011. When the censored proportion was less than 30%, censored Poisson regression model had the average relative error less than 20%. Since censored Poisson regression model directly dealt with censored count data without imputing or deleting, so it kept all the information of the censored count datasets.

The relative errors for censored Poisson regression model were less than 1% even when the censored proportions were greater than 50% in the simulated censored count datasets, while in the real-world censored count datasets, the relative error dramatically increased (from 16% to 80%) when the censored proportion was larger than 30%. This finding indicates a weakness of censored Poisson regression model. Censored Poisson regression model is restricted by the distribution assumption of the censored count datasets. In simulated censored count datasets, the outcomes followed a Poisson distribution, but in the real-world censored count datasets, the outcomes cannot reach the ideal scenario. Thus, the relative errors of censored Poisson regression fluctuated when the censored proportion exceeded 30%. Second, the censored Poisson regression had the censored proportion limitations. When the censored proportion exceeds some value (in this paper, the censored

proportion was greater than 37.88%), we cannot rely on the estimations results of censored Poisson regression.

Despite the limitations, this study has some strengths. First, we found that censored Poisson regression model had stable performances under different censored proportions, especially under the censored proportion of 30% in real-world data, which provided a strong method to deal with censored count datasets in spatial analysis. Second, we improved computational efficiency of Censored Poisson regression model by using INLA algorithm rather than the other typically used algorithms (i.e., MCMC and Newton Raphson) [10] for parameter estimates.

In conclusion, censored Poisson regression model had stable and accurate estimations dealing with censored count data under different censored proportions, especially under the censored proportion less than 30%. In the future, we could compare more different methods to deal with censored count data and compare the results of parameter estimations, in order to find a most accurate and efficient method to deal with censored datasets.

Tables for Journal Article #1

| | 7.9% | | 33.6% | | 54.1% | |
|------------------------|-----------------------|--------|-----------------------|--------|-----------------------|--------|
| | Mean (Relative Error) | S.D | Mean (Relative Error) | S.D. | Mean (Relative Error) | S.D. |
| b1 | 2.0031 (0.15%) | 0.0501 | 2.0071 (0.35%) | 0.0577 | 2.0106 (0.53%) | 0.0474 |
| b2 | 2.9979 (0.07%) | 0.0998 | 3.0009 (0.30%) | 0.1103 | 3.0092 (0.31%) | 0.0944 |
| b3 | 3.9842 (0.40%) | 0.0502 | 4.0011 (0.30%) | 0.0577 | 4.0155 (0.39%) | 0.0474 |
| Average Relative Error | 0.21% | | 0.32% | | 0.41% | |

Table 1. Simulation results for censored Poisson regression model under different censored proportions

| | Poisson Regression Model | | 4.56% | | 8.08% | | 12.96% | | 20.41% | |
|------------------------|--------------------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | 0.0003 | 0.0015 | 0.0003 | 0.0015 | 0.0004 | 0.0016 | 0.0004 | 0.0016 |
| lag1heat | 0.0002 | 0.0019 | 0.0001 | 0.0019 | 0.0001 | 0.0019 | 0.0001 | 0.0019 | 0.0002 | 0.0019 |
| lag2heat | -0.0015 | 0.0019 | -0.0014 | 0.0019 | -0.0013 | 0.0019 | -0.0014 | 0.0019 | -0.0015 | 0.0019 |
| lag3heat | -0.0010 | 0.0016 | -0.0012 | 0.0015 | -0.0012 | 0.0015 | -0.0013 | 0.0015 | -0.0010 | 0.0016 |
| time1 | -1.3970 | 0.2980 | -1.4238 | 0.2978 | -1.4155 | 0.2982 | -1.3705 | 0.2989 | -1.4219 | 0.3011 |
| time2 | 1.1800 | 0.5821 | 1.2375 | 0.5818 | 1.2385 | 0.5825 | 1.2258 | 0.5839 | 1.2624 | 0.5876 |
| time3 | 0.0246 | 0.2961 | 0.0189 | 0.2951 | 0.0208 | 0.2955 | -0.0152 | 0.2963 | -0.0365 | 0.2986 |
| Average Relative Error | | | 20.99% | | 21.07% | | 16.21% | | 8.63% | |

Table 2. Case study results for censored Poisson regression model under different censored proportions

| | Poisson Regression Model | | 28.07% | | 37.88% | | 48.65% | | 59.33% | |
|------------------------|--------------------------|--------|---------|--------|---------|--------|---------|--------|---------|--------|
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | 0.0000 | 0.0016 | -0.0005 | 0.0016 | -0.0002 | 0.0016 | -0.0006 | 0.0017 |
| lag1heat | 0.0002 | 0.0019 | 0.0002 | 0.0020 | 0.0003 | 0.0020 | -0.0003 | 0.0020 | 0.0006 | 0.0021 |
| lag2heat | -0.0015 | 0.0019 | -0.0017 | 0.0019 | -0.0016 | 0.0020 | -0.0017 | 0.0020 | -0.0021 | 0.0021 |
| lag3heat | -0.0010 | 0.0016 | -0.0007 | 0.0016 | -0.0008 | 0.0016 | -0.0008 | 0.0016 | -0.0010 | 0.0017 |
| time1 | -1.3970 | 0.2980 | -1.4156 | 0.3035 | -1.5095 | 0.3087 | -1.3711 | 0.3168 | -1.3410 | 0.3294 |
| time2 | 1.1800 | 0.5821 | 1.2037 | 0.592 | 1.0145 | 0.6022 | 0.9357 | 0.6170 | 0.9130 | 0.6398 |
| time3 | 0.0246 | 0.2961 | -0.0834 | 0.3013 | -0.1012 | 0.3077 | -0.0933 | 0.3152 | -0.0260 | 0.3276 |
| Average Relative Error | | | 78.91% | | 154.38% | | 151.21% | | 98.90% | |

Table 3. Case study results for censored Poisson regression model under different censored proportions

Figures for Journal Article #1

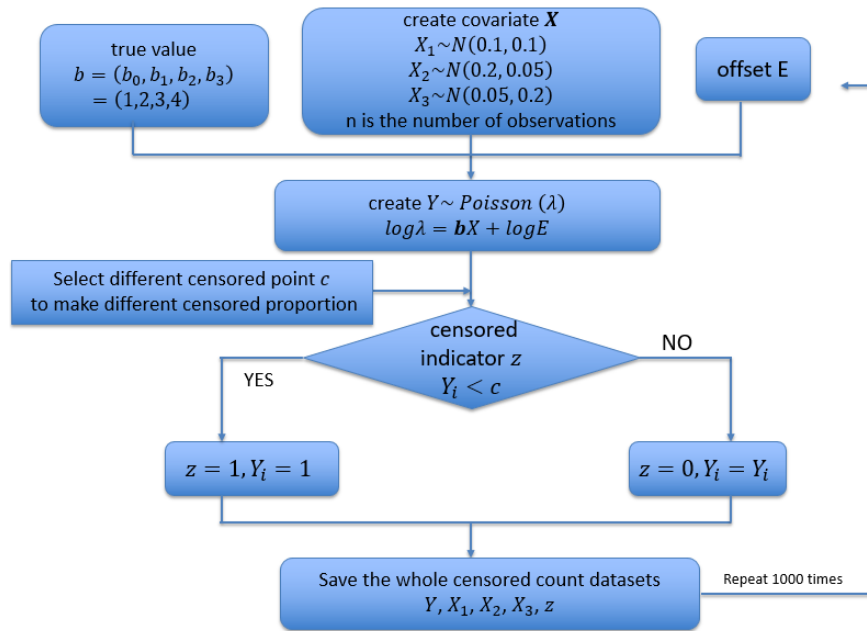


Figure 1. Flow Chart for Simulating Censored Dataset

REFERENCES

- [1] data.illinois.gov. IDPH STD Illinois By County Rank. 2017.
- [2] Prevention CfDCa. Children Blood Lead Sheet. 2008.
- [3] Hyun S, Newell, James. Survival Probabilities with and without the Use of Censored Failure Times. USC Upstate Undergraduate Research Jornal. 2011;IV:35-8.
- [4] Whitehead NS, Leiker R. Case management protocol and declining blood lead concentrations among children. Prev Chronic Dis [serial online]. 2007.
- [5] Dignam TA, Lojo J, Meyer PA. Reduction of elevated blood lead levels in children in North Carolina and Vermont, 1996–1999. Environmental health perspectives. 2008;116:981.
- [6] Frome E. PREG: A computer program for Poisson regression analysis. Oak Ridge Associated Universities, Inc., TN (USA); 1981.
- [7] Mahmoud MM, Alderiny MM. On estimating parameters of censored generalized Poisson regression model. Applied Mathematical Sciences. 2010;4:623-35.
- [8] Terza JV. A Tobit-type estimator for the censored Poisson regression model. Economics Letters. 1985;18:361-5.
- [9] Veleba J. Stability algorithms for Newton-Raphson method in load flow analysis. 2010.
- [10] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology). 2009;71:319-92.
- [11] Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology. 2013;4:33-49.
- [12] Blangiardo M, Cameletti M. Spatial and spatio-temporal Bayesian models with R-INLA: John Wiley & Sons; 2015.

[13] Information NCE. Integrated Surface Database. 2016.

JOURNAL ARTICLE #2

Performance of Multiple Imputation, Small Area Estimation, and Censored Poisson Regression Model When Handling Censored Count Data

Journal of Biometrics and Biostatistics

Yu X¹, Chan W¹, Swint JM¹, Zhang K¹, Chien LC².

¹ *School of Public Health, University of Texas Health Science Center at Houston, TX, USA*

² *Community Health Sciences, University of Nevada, Las Vegas, NV, USA*

Abstract

Background: Count data are typically used to report frequency statistics of diverse health outcomes. However, some data are marked on purpose to avoid leaking information that could be used to identify individuals when population sizes are small. The situation hinders the further use from those data in public health research. Thus, an accurate and efficient method for dealing with censored count data is needed.

Methods: We applied three methods 1) multiple imputation (MI) method; 2) small area estimation (SAE) method; 3) censored Poisson regression model (CPRM) method on both simulated censored count datasets and real-world censored count datasets to the association between heat wave and cardiovascular diseases using hospital admission data from Harris County, Texas, from 2006 to 2011 under different censored proportions. We calculated the relative errors, depicted graph the bar charts, and recorded the computational time to compare the accuracy and efficiency of these three methods.

Results: In the simulation study, we found that CPRM had the lowest relative error and MI method had the shortest computational time. In the case study, when the censored proportion was less than 30%, CPRM had the best accuracy, but when the censored proportion was greater than 30% but less than 40%, SAE yielded the most accurate parameter estimates of all the methods.

Conclusions: By balancing the computational time and estimation accuracy, CPRM is the most appropriate method to deal with censored count data.

Key words: censored count data, censored Poisson regression model, multiple imputation (MI), small area estimation (SAE)

Introduction

In public health, count data sets are commonly used to report statistics related to emerging or existing health problems, and play a significant role in biostatistics. In fact, most health reports published by the U.S. Centers for Disease Control and Prevention (CDC) are based on count data. For example, the CDC's Birth Defects Countries and Organizations United for Neural Tube Defects Prevention initiative reports that 3,000 pregnancies in the United States are affected by neural tube defects each year [1] and estimates that folic acid fortification may reduce the prevalence of neural tube defects by 50% or more. In this case, count data sets are being used to help prevent neural tube defects, and associated morbidity and mortality rates. In addition, the CDC reports count data for cases of Lyme disease by county, state, and year, which allows the prevalence of Lyme disease to be analyzed geographically and temporally. Data show that cases of Lyme disease are concentrated in the Northeast and Upper Midwest regions of the United States, which enables targeting of prevention efforts, i.e., those states with a higher prevalence of Lyme disease can dedicate more resources to prevent it [2].

One situation commonly exists in the count data in public health, especially for spatial analysis, is the censored situation in the count data sets. In spatial analysis, the majority of censored situations are left censored, which means that a data point is below a certain value, but it is unknown by how much [3]. For example, in the CDC's Childhood

Lead Poisoning Prevention Program surveillance database, the number of children with elevated blood lead levels in each county is censored by 5 or less [4].

In previous studies, censored data were considered as missing, however, censored data are not actually missing, but just intentionally masked, which may lead to biased results. For example, a study applying the Cox regression model on the reduction of blood lead levels shows that the results with the censored data are less biased than those without the censored data [5, 6]. Even in the simple linear regression model, removing the censored data could cause the bias in the estimation [7]. Furthermore, in a study of hypertension treatment using the generalized hierarchical multivariate conditional autoregressive model, when 24 censored data points were considered missing, 80% of patients completed the protocol with effective control of hypertension and no side effects; however, when the censored data were accounted for, the percentage of patients was 44% instead of 80% [8, 9].

In this study, we want to apply three methods dealing with censored count data and find the most accurate the efficient method. One way to deal with the censored count data is the multiple imputation method. The mission of multiple imputation is to create a complete data set, then statistical models could be used as usual. Imputation method was first developed to deal with missing data problem in the 1980s [10,11]. Compared with the mean imputation method multiple imputation methods had a better performance and was applied in public health research data [12]. In addition, the multiple imputation method has been developed to solve advanced missing data problems, such as nonparametric multiple imputation, multilevel multiple imputation, and so on [13-17]. In this paper, we applied multiple imputation method to censored count data instead of missing data.

A second way to deal with the censored part of the count data, which is frequently used in the spatial analysis, is the small area estimation method. Small area estimation is used when traditional demographic sample surveys designed for national estimates do not provide large enough samples to produce reliable direct estimates for small areas such as counties. The statistical methods using for small area estimation developed dramatically in recent years. Bayesian unit-level model estimates the prevalence of diabetes at each county in the U.S. [18]. This model can analyze the prevalence for each county level considering different independent variable layers. For example, the method could estimate the prevalence for the specific county, specific gender and specific age group. Then the model was extended by changing the distribution of the prevalence from Poisson to Binomial, and estimated the diabetes incidence for each county by different layers [19]. A research on chronic obstructive pulmonary disease (COPD) developed a multilevel logistic model to generate small-area estimates of the prevalence of COPD in different geographic unites [20]. An analysis of the drinking pattern used a spatiotemporal model to estimate county-level alcohol use prevalence in the U.S [21]. This method considered spatial and temporal information as covariates to improve the predictions of all areas, as long as the area with limited sample sizes.

The third way to deal with the censored count data is censored Poisson regression model. The model was first developed by Terza (1985) and Newton-Raphson algorithm was used to estimate unknown parameters [22]. Famoye and Wang (2004) expanded the censored Poisson regression model to the censored generalized Poisson regression model to handle censored data with over-dispersion or under-dispersion [23]. They used an iterative algorithm to get the maximum likelihood estimators, but they did not specify the iterative algorithm

used. Mahmoud (2010) also developed a censored generalized Poisson regression model using a method similar to Famoye and Wang's method by adopting the Newton-Raphson algorithm [24]. In this paper, we used integrated nested Laplace approximation (INLA) to estimate unknown parameters in censored Poisson regression to deal with censored count data.

In this paper, we compared the estimation results and the computational time of three different methods, which are MI, SAE and CPRM dealing with both simulation censored count datasets and real-world censored count datasets under different censored proportions, in order to find the most accurate and efficient method to prevent diseases by better determine the risk factors and know the patterns of disease.

Methods

Statistical Methods

Multiple Imputation

The multiple imputation method uses a set of values replacing the missing values instead of using a single value for each missing datum. The “mi” or “mice” R packages are usually used to perform multiple imputation. In these R packages, Bayesian models were used to impute the data more precisely by giving multiple values than single values. Based on different properties for different data type, R packages provided different functions, i.e., if the data were binary, *mi.binary()* function was used; if the data were count data, *mi.count()* function was used.

We used multiple imputation to impute the censored count instead of missing data. We applied multiple imputation to simulated censored count dataset and real-world count dataset. We performed the following steps to impute the censored count data:

Step 1: Simulated the censored values independently using the estimated mean vector and covariance matrix. Censored observation (observation without values) was represented by $Y_{i(cen)}$, While the observation with value was represented by $Y_{i(obs)}$. An imputed $Y_{i(cen)}$ was drawn from a conditional distribution $Y_{i(cen)}|Y_{i(obs)}$.

Step 2: Simulated the posterior population mean vector and covariance matrix X from the complete sample estimates using a non-informative prior, which was built in R packages.

Step 3: Repeated step 1 and step 2 for 5 times as recommended by Rubin (1987) [58], which was built in R packages.

Step 4: Used the Poisson regression model with covariance matrix X and Y to impute the censored observation $Y_{i(cen)}$.

Step 5: Averaged the values and the standard errors of the parameter estimations across the censored value samples in order to obtain a single point estimate.

R packages “*mi*” and “*mice*” were applied to generate the values for the censored outcomes in order to produce the complete datasets. After generating the complete datasets, we used the Poisson regression model to estimate the coefficients of covariate matrix X.

Small Area Estimation

The small area estimation method is a method commonly used in spatial analysis to estimate the unknown value for small counties. The idea of small area estimation uses the known observations of an outcome to estimate the unknown observations by the stratified demographic variables (e.g., age, gender, and race). We used small area estimation to estimate the censored count data by stratified demographic variables.

We assumed Y_{ijkc} as the count of an outcome variable (e.g. the number of cases) at age group i , race j , gender k in county c , which follows a Poisson distribution with a mean of λ_{ijkc} . Thus, the model was specified as follows,

$$\log(\lambda_{ijkc}) = \alpha + \beta_{1i} + \beta_{2j} + \beta_{3k} + f_{spat}(c) + \log(n_{ijkc}), \quad (1)$$

where β_{1i} , β_{2j} , and β_{3k} were fix effects for age, race, and gender, respectively. The spatial function $f_{spat}(c)$ was Markov random fields following an intrinsic conditional autoregressive prior [61]. The last term $\log(n_{ijkc})$ was an offset corresponding to the logarithm of the at-risk population index by i, j, k , and c (the total number of individuals corresponding to Y_{ijkc}). We applied INLA algorithm described in Cadwell et al. [45] to estimate the censored count data.

We defined N_{ijkc} and Y_{ijkc} in equation (1) as age-race-gender-county specific at-risk population and those with the specific outcome, respectively. Thus, we derived Z_{ijkc} , the number of unobserved individuals (the censored cases) with the specific outcome, indexed by age, race, gender, and county straightforwardly. The sum of the observed and unobserved cases, $Y_{ijkc} + Z_{ijkc}$, was the total count of the outcome, where

$$Z_{ijkc}|Y_{ijkc}, n_{ijkc}, N_{ijkc} \sim \text{Poisson}(\mu_{ijkc}).$$

The parameter Y_{ijkc} was defined as

$$Y_{ijkc} = \left(\frac{\hat{\lambda}_{ijkc}}{n_{ijkc}} \right) \times (N_{ijkc} - n_{ijkc}) = \frac{\exp(\hat{\alpha} + \hat{\beta}_{1i} + \hat{\beta}_{2j} + \hat{\beta}_{3k} + \hat{f}_{spat}(c))}{n_{ijkc}} \times (N_{ijkc} - n_{ijkc}).$$

We applied the small area estimation method to estimate the censored count data, and then analyzed the data with the Poisson regression model to estimate the coefficient of all the covariates.

Simulation Study and Case Study

The procedure of simulating the censored count datasets was shown in the previous section. For multiple imputation method, we applied R packages “mi” and “mice” to perform the multiple imputation on the simulated censored count datasets. Function *mice()* was applied to impute data, then *pool()* function was used to take average of imputed data. After generating the complete dataset, we applied Poisson regression model to the complete data and estimated the parameters for covariate matrix, whose true values were b_1, b_2, b_3 , and b_4 . We calculated the relative error for each estimation, then we saved the estimation results for accuracy comparison.

For small area estimation methods, we used the existing individual demographic variables to estimate the unknown outcome observations. We completed the following steps:

Step 1: Used the demographic variables to estimate the censored count data. We used stratified demographic variables, such as age group and race, to build up the model to estimate the censored count data. We used R packages “*inla*” to do the small area estimation. We used function *inla()* to estimate the full dataset of outcome dataset, then we extracted the fitted value from demographic model. According to the indicator variable Z , we identified the censored outcome data.

Step 2: Completed censored outcome data. We tried the following two ways to complete the censored part of the dataset. First, we used the estimated Y , even if it exceeded the censored point. For example, when the censored point was 7, the values lower than 7 were censored. When the fitted outcome was 8, we still used 8 to impute the censored value, even though 8 was larger than 7. Second, when the estimated Y exceeded the censored point, we used the censored point instead of the estimated Y . For example, when the censored point was 7, but the imputed value was 8 (larger than 7), we used 7 (the censored point) to impute the values instead of 8 (the fitted value).

Step 3: Estimated the coefficients for covariate matrix. We used Poisson regression model to estimate the coefficient of independent variables. We used *glm()* function to manipulate Poisson regression model and estimated the parameters, whose true values were b_1 , b_2 , b_3 , and b_4 . We repeated 1000 datasets to calculate the average of mean and standard deviation for each coefficient.

For censored Poisson regression model, we used R package “INLA” and directly applied a censored Poisson regression model to the existing censored count datasets. We used

INLA algorithm to estimate the coefficients (b_1 , b_2 , b_3 , and b_4), whose true values were 1, 2, 3, and 4, respectively. Then we compared the accuracy of this method using relative error, and plotted bar charts to depict the estimation results.

Case study datasets are the same as the methods describe, we first complied the individual level admission data to county level data. Second, we chose different censored point to make different censored proportion. We compared the results from the following 5 models: (1) Poisson regression model under complete datasets; (2) Poisson regression model directly applied to censored count data (censored data with Poisson regression model, short as CDPRM); (3) multiple imputation method dealing with censored count data (MI method); (4) small area estimation method dealing with censored count data (SAE); (5) censored Poisson regression model (CPRM).

We plotted bar plots to show the value of different estimation for small area estimation under different censored proportions for both simulation and case study. We compared the accuracy of results of parameter estimations by calculating the relative errors. We used R version 3.2.3 for all the statistical analysis.

Results

Simulation Study

Tables 1 presents the simulation results of the multiple imputation method, small area estimation method, and censored Poisson regression method to compare the accuracy of the estimation of parameters under different proportions of censored count datasets. We chose

the censored point equal to 7, 10, and 12, resulting in censored proportion by 7.9%, 33.6%, and 54.1%, respectively.

When the censored proportion was 7.9%, the average relative error of MI, SAE, and CPRM were 9.13%, 4.93%, and 0.21%, respectively (shown in Table 2). MI had the largest relative error in all estimated parameters, while CPRM had the smallest relative error. Across the three methods, the standard deviations were similar for b_0 to b_2 . For b_3 , however, the standard deviation for MI was larger than that for SAE and CPRM. In terms of computational time among the three methods, SAE had longest time (124.11s, data not shown in the table), whereas MI had the shortest time (19.30s, data not shown in the table). The average computational time for MI, SAE, and CPRM for different censored proportions was shown in table 1.

When the censored proportion was 33.6%, the average relative error of MI, SAE, and CPRM were 33.62%, 6.01%, and 0.32%, respectively (shown in Table 2). When the censored proportion was 54.1%. The average relative error of MI, SAE, and CPRM were 52.88%, 33.00%, and 0.41%, respectively (shown in Table 2). Across different censored proportions, MI had the largest relative error in all estimated parameters, while CPRM had the smallest relative error. Even under the censored proportion of 50%, the relative errors for CPRM were still less than 1%. Across the three methods, the standard deviations were similar for b_0 to b_2 . For b_3 , however, the standard deviation for MI was larger than that for SAE and CPRM. In terms of computational time among the three methods across different censored proportions, SAE had longest time, where MI had the shortest time.

To facilitate the comparison of parameter estimation across methods, Appendix Figure 1 shows the bar plot for all three methods under different censored proportions. Appendix Figure 2 shows the bar plot for the standard deviation for parameter estimations across three methods under different censored proportions.

Case Study

Figures 1-8 present the case study results of the following five methods: (1) Poisson regression model (2) censored data PRM (3) MI (4) SAE (5) CPRM to compare the accuracy of the parameter estimations under different censored proportions. We chose the censored points equal to 7, 8, 9, 10, 11, 12, 13, and 14, resulting in censored proportion by 4.56%, 8.08%, 12.96%, 20.41%, 28.07%, 37.88%, 48.65%, and 59.33%, respectively. The details of the results shows in Appendix Table 1 – Table 8.

Figure 1 presents the results when the censored proportion was 4.56%. The average relative error of censored data PRM, MI, SAE, and CPRM were 77.39%, 126.58%, 86.58%, and 20.99%, respectively (shown in Table 3). MI had the largest relative error in all estimated parameters, while CPRM had the smallest relative error. Because the estimation of the parameters were quite small (i.e., the coefficient for Max Heat is 0.0001) in our complete data, the relative error was large under this situation (when the true value is small). However, the relative error still could provide a standard for judging the accuracy of different methods, at the same time, we need to consider the sign of the estimation, since relative error is an absolute value. The signs of the coefficient estimation for MI and CPRM were the same as that of the true values. The sign of lag1heat for censored data PRM was the opposite from the

true value and the sign of lag1heat and time3 were opposite from the true value. Across the four methods, the standard deviations were similar for Max Heat, lag1heat, lag2heat, lag3heat, time1, and time2 with the standard deviations of true values. For time3, however, the standard deviation for censored data CPM was smaller than the other methods, which have the similar results with the true value. Figure 9 shows the bar plot for standard deviation for Max Heat under different proportions, the other standard deviation shows in the Appendix Table 1 - 8. The computational time were 0.02s, 1.32s, 10.42s, and 5.79s for censored data PRM, MI, SAE, and CPRM, respectively. The computational time for each method remained almost the same under different censored proportion.

For results shown in Figure 2–7, we found the following results: when the censored proportion was less than 10%, the relative errors for censored data PRM, MI were around 200%, which were not appropriate for handling the censored count data anymore; when the censored proportion was greater than 10% but less than 30%, CPRM had the lowest relative errors; when the censored proportion was greater than 30% but less than 45%, SAE had the lowest relative errors. When the censored proportion exceeds 45%, all the methods had large relative errors and the sign of the parameters turned to the opposite, which means the estimation results under the censored proportion of 45% were not reasonable. Figure 10 shows the relative errors for different methods under different censored proportions. We found that for overall proportions, CPRM had relative low relative errors. In addition, all relative errors for different methods were fluctuated. The standard deviations for all the coefficients across all the methods were similar, except the one for time3 using censored data PRM, which was smaller than the others. The computational time did not change with the

censored proportion. Censored data PRM had the shortest computational time, while SAE had the longest computational time.

Discussion

We evaluated the performance of multiple imputation (MI), small area estimation method (SAE) and censored Poisson regression method (CPRM) in dealing with censored count data. Using both simulated and real-world censored count data, the censored Poisson regression method performed the best, yielding the most accurate parameter estimates with an efficient computational time.

In our simulation study, the simulated censored count data represented an ideal scenario, in which the outcome variable followed a Poisson distribution with a controlled mean and variance (the mean was equal to the variance). Under this condition, the censored Poisson regression model performed exceptionally well (relative error less than 1%) even when the censored proportion was greater than 50% censored proportion. SAE was restricted by the stratified demographic variables. When the censored proportion was less than 40%, the relative errors of SAE were fluctuated at around 5%. MI was only applicable when the censored proportion was less than 10%, when MI was applied, its performance was worse than that of SAE and CPRM. The computational time of MI was the shortest dealing with large censored datasets. Overall, as the censored proportion increased. The relative errors also increased. Specifically, when the censored proportion increased from 33.6% to 54.1%, the relative errors of MI and SAE dramatically increased, but the relative error of CPRM

remained the same. Thus, CPRM offers the best performance for dealing with censored count datasets under ideal conditions.

In our case study, we assessed the association between heat wave temperature and hospitalization due to cardiovascular diseases in Harris County, Texas, from 2006 to 2011. We found that each method (censored data PRM, MI, SAE, CPRM) had both strengths and weakness under different censored proportions.

Directly applying a Poisson regression model to censored count dataset (censored data PRM) represented the typical method used to handle censored count data. When the censored proportion was less than 5%, the censored data PRM was more accurate than MI and SAE, but less accurate than CPRM. In addition, censored data PRM had the shortest computational time of all methods. When dealing with large censored count datasets with a censored proportion less than 5%, censored data PRM is a strong method that balances accuracy with computational efficiency. However, when the censored proportion is greater than 5%, the accuracy of the censored data PRM decreases, because it does not account for the censored counts, which are not actually missing values and thus need to be accounted for.

Furthermore, under all censored proportions in our case study, even less than 5%, MI yielded inaccurate parameter estimates. The results were probably caused by the following reasons. First, we could not control the imputed values within the censored interval. For example, if the values were less than 10, the values were censored. However, the imputed values could be greater than 10, which exceed the censored interval. Second, we could not control the distribution of the imputed outcomes. Since we assumed that the outcomes

followed Poisson distribution where the mean is equal to the variance, the imputed outcomes cannot meet the distribution assumption. Therefore, although the computational time of MI was short, the accuracy for MI was low under each censored proportion.

Moreover, when the censored proportion was greater than 30% but less than 40%, SAE yielded the most accurate parameter estimates of all the methods. SAE was the only method that considered the demographic information when estimating the censored count outcomes. This use of demographic information is both a strength and a weakness of SAE. According to Barker et al. [46], the more stratified the demographic levels are, the better the performance of SAE. For example, the estimations for data with more stratified demographic variables, i.e., gender and race (the number of male white, female white, male black, and female black) were better than the estimations for data with fewer stratified demographic variable i.e. gender only (the number of male and female). The weakness of SAE co-occurred with the strengths. First, the more stratified demographic levels included, the slower the computational time. If the sample size was extremely large, the computational time would be extremely long and may not be obtainable, because it exceeds the computational ability. Second, for most real-world datasets, we cannot extract all the demographic information we need, which also restricts the accuracy of SAE method.

Lastly, when the censored proportion was less than 30%, CPRM had the best performance of all methods by balancing accuracy with computational efficiency. The relative error of CPRM remained stable (around 16%) and was the lowest among all the methods. CPRM directly dealt with censored count data without imputing or deleting, so it kept all the information of the censored count datasets. The relative errors for CPRM were

less than 1% even when the censored proportions were greater than 50% in the simulated censored count datasets, while in the real-world censored count datasets, the relative error dramatically increased (from 16% to 80%) when the censored proportion was larger than 30%. This finding indicates a weakness of CPRM. CPRM is restricted by the distribution assumption of the censored count datasets. In simulated censored count datasets, the outcomes followed a Poisson distribution, but in the real-world censored count datasets, the outcomes cannot reach the ideal scenario. Thus, the relative errors of censored Poisson regression fluctuated when the censored proportion exceeded 30%.

This study has some important limitations. First, we did not identify trends or pattern in the fluctuation of relative errors as the censored proportion increased across all methods. Second, we used a real-world censored count datasets may generate results than other real-world censored count datasets. Thus, we only provided a reference instead of a standard for determining the best performance under different specific censored proportions. In the future, these methods could be applied to additional real-world censored count datasets to identify a relatively precise standard for different censored proportions. Third, we used censored count datasets that met Poisson distribution assumption, which provided better simulation results than real-world results. In the future, more accurate models to deal with censored count data with less distribution assumptions could be developed.

Despite the limitations, this study has some strengths. First, to our knowledge, this study is the first to compare all the methods dealing with censored count data considering both estimation accuracy and computational efficiency simultaneously. Thus, we identified the best method under different censored proportions, which provides a means for

determining the best method to use based on the size of the censored count dataset and of the censored proportion. Second, we improved computational efficiency of CPRM by using INLA algorithm rather than the other typically used algorithms (i.e., MCMC and Newton Raphson) for parameter estimates [25].

In conclusion, considering the balance of the estimation accuracy with computational time, the censored Poisson regression model is the best method for dealing with censored count datasets under different censored proportions, especially when the censored proportions were less than 30%. However, when the censored proportions were greater than 30% and stratified demographic data can be collected, the small area estimation method performed well, but it had longer computational time. MI method had the shortest computational time, but only applicable when the sample size was large and the censored proportion was low (less than 5%). Future research is need to identify trends or patterns in the fluctuation of relative errors as the censored proportion increase. In addition, more accurate models to deal with censored count data with less distribution assumptions could be developed.

Tables for Journal Article #2

| Censored Proportion | Parameters | MI (18.27s) | | SAE ^a (119.87s) | | CPRM (61.02s) | |
|---------------------|------------|-----------------------|--------|----------------------------|--------|-----------------------|--------|
| | | Mean (Relative Error) | S.D. | Mean (Relative Error) | S.D. | Mean (Relative Error) | S.D. |
| 7.9% | b1 | 1.8189 (9.06%) | 0.0473 | 1.9055 (4.73%) | 0.0467 | 2.0031 (0.15%) | 0.0501 |
| | b2 | 2.7330 (8.90%) | 0.0941 | 2.8501 (5.00%) | 0.093 | 2.9979 (0.07%) | 0.0998 |
| | b3 | 3.6221 (9.45%) | 0.1879 | 3.7976 (5.07%) | 0.0467 | 3.9842 (0.40%) | 0.0502 |
| 33.6% | b1 | 1.3286 (33.57%) | 0.0472 | 2.1259 (6.29%) | 0.0466 | 2.0071 (0.35%) | 0.0577 |
| | b2 | 2.0051 (33.16%) | 0.0941 | 3.1769 (5.89%) | 0.0928 | 3.0009 (0.30%) | 0.1103 |
| | b3 | 2.6353 (34.12%) | 0.1838 | 4.2341 (5.85%) | 0.0466 | 4.0011 (0.30%) | 0.0577 |
| 54.1% | b1 | 0.9394 (53.03%) | 0.0501 | 2.6655 (33.27%) | 0.0479 | 2.0106 (0.53%) | 0.0577 |
| | b2 | 1.4156 (52.81%) | 0.0998 | 3.9911 (33.04%) | 0.0954 | 3.0092 (0.31%) | 0.1103 |
| | b3 | 1.8882 (52.79%) | 0.0502 | 5.3078 (32.69%) | 0.0479 | 4.0155 (0.39%) | 0.0577 |

Table 1. Parameter estimation results across three methods under different censored proportions

^a Results are calculated using the second method for SAE described in Method Section. Results of estimation for the two ways of SAE did not differ significantly.

| Censored Proportion | MI | SAE | CPRM |
|---------------------|--------|--------|-------|
| 7.9% | 9.13% | 4.93% | 0.21% |
| 33.6% | 33.62% | 6.01% | 0.32% |
| 54.1% | 52.88% | 33.00% | 0.41% |

Table 2. Average relative errors for different methods under different censored proportions

| Censored Proportion | censored data PRM | MI | SAE | CPRM |
|---------------------|-------------------|---------|---------|---------|
| 4.56% | 77.39% | 126.58% | 86.58% | 20.99% |
| 8.08% | 195.33% | 206.67% | 118.45% | 21.07% |
| 12.96% | 322.81% | 327.18% | 103.72% | 16.21% |
| 20.41% | 501.14% | 503.46% | 879.45% | 8.62% |
| 28.07% | 269.36% | 589.27% | 36.17% | 78.90% |
| 37.88% | 174.78% | 117.14% | 49.20% | 154.38% |
| 48.65% | 132.98% | 77.08% | 23.47% | 151.21% |
| 59.33% | 265.46% | 318.65% | 600% | 98.90% |

Table 3. Average relative errors for all methods under different censored proportions

Figures for Journal Article #2

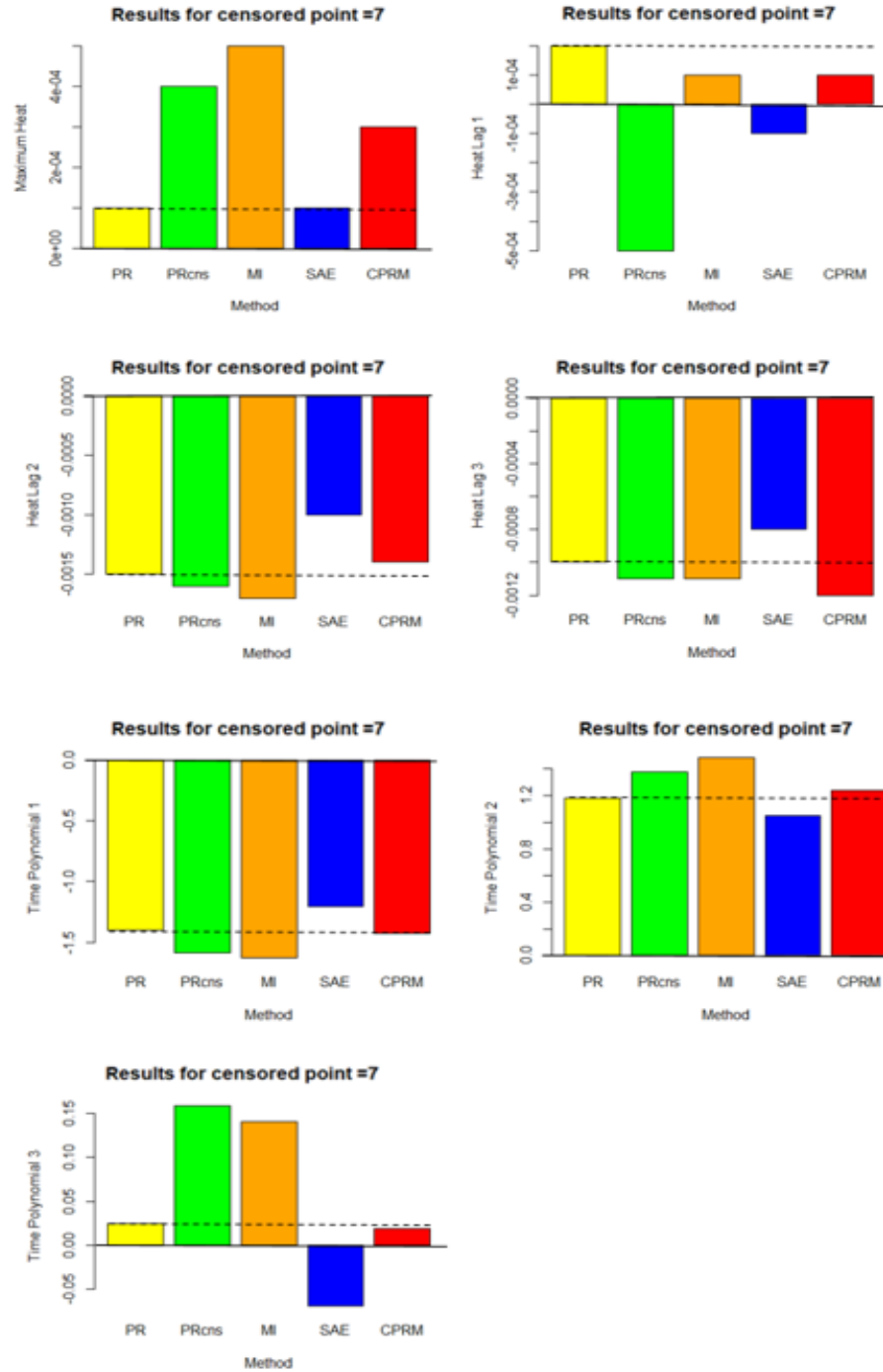


Figure 1. Bar plot for estimations under censored point = 7 (censored proportion = 4.56%)

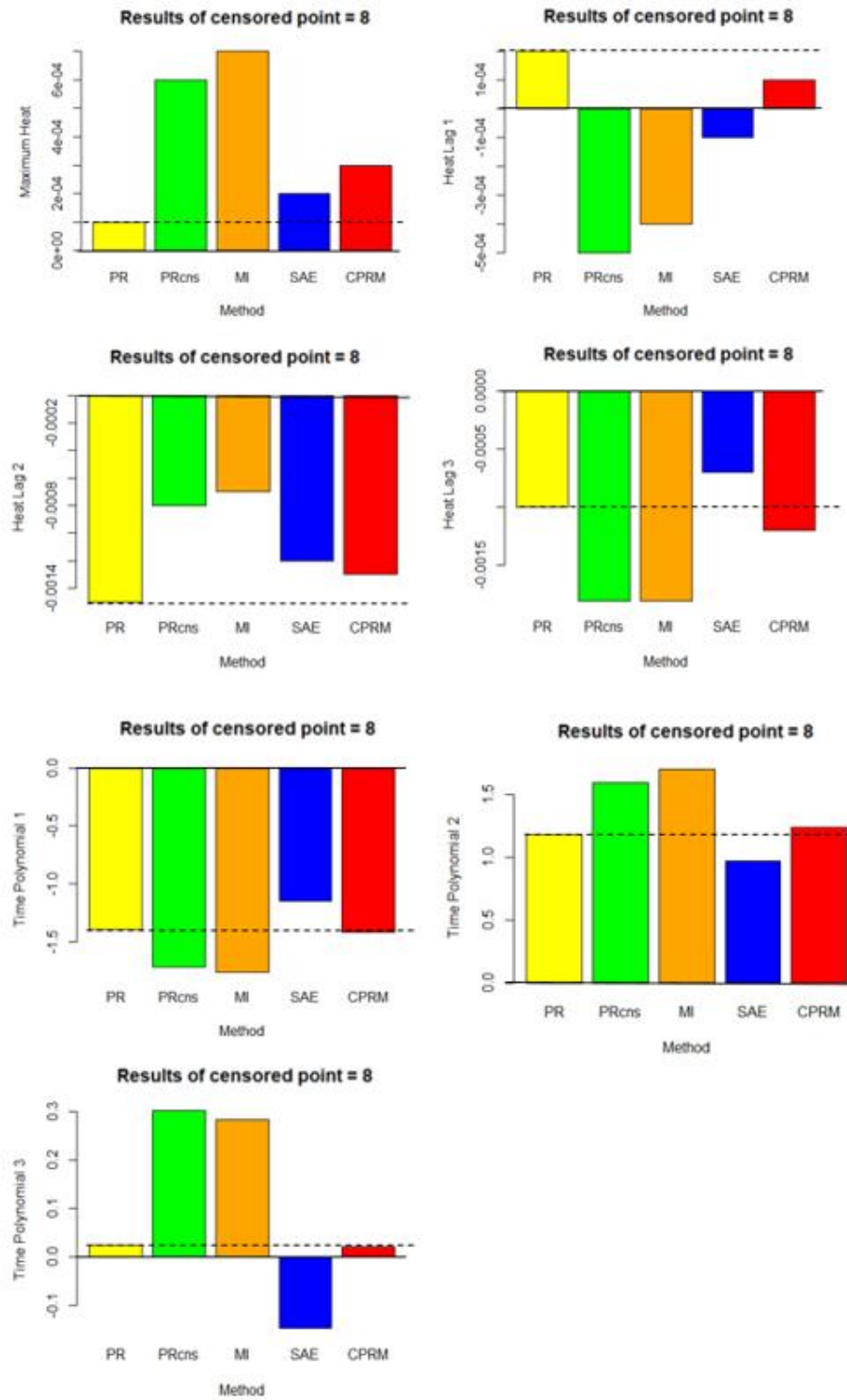


Figure 2. Bar plot for estimations under censored point = 8 (censored proportion = 8.08%)

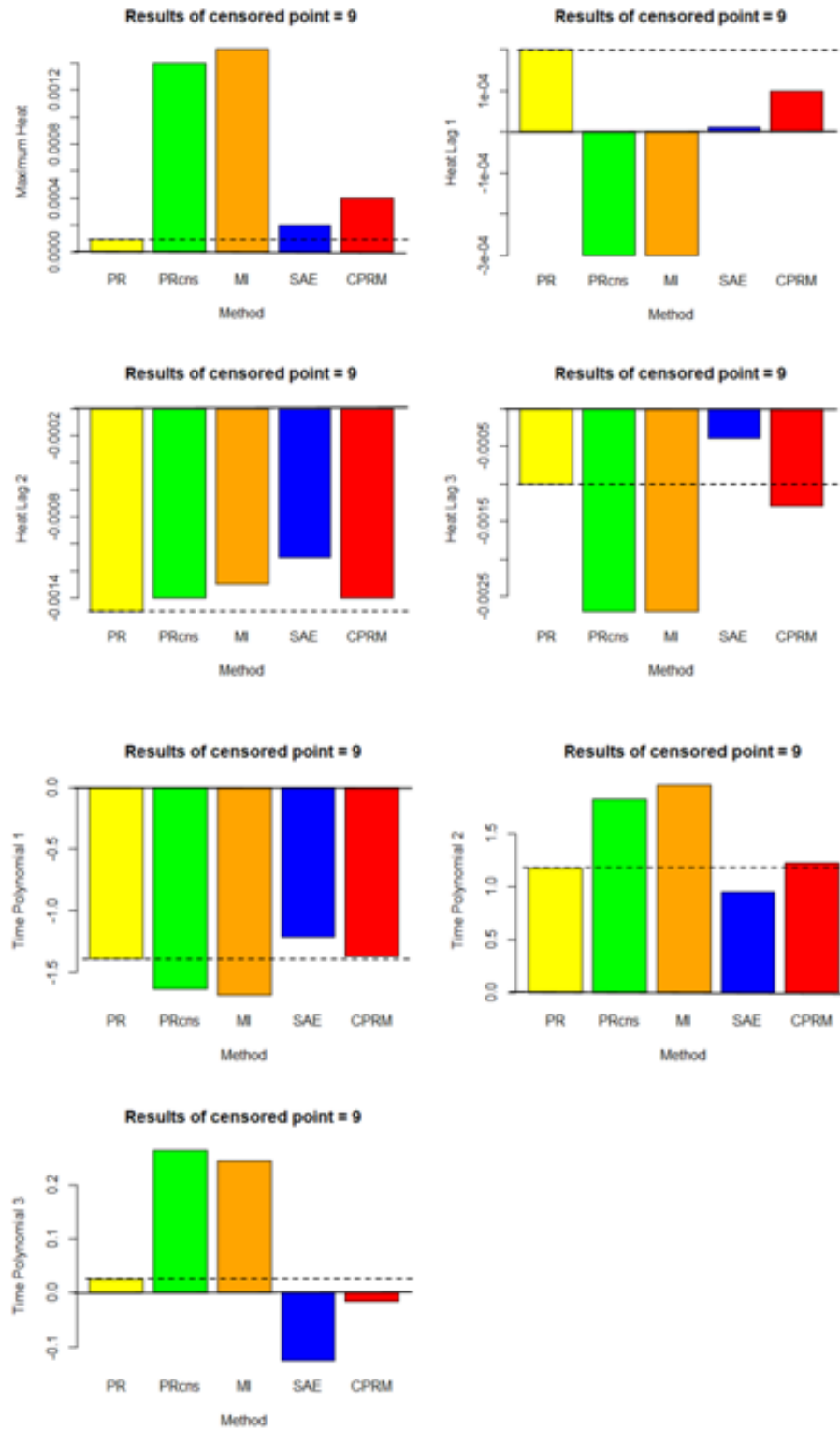


Figure 3. Bar plot for estimations under censored point = 9 (censored proportion = 12.96%)

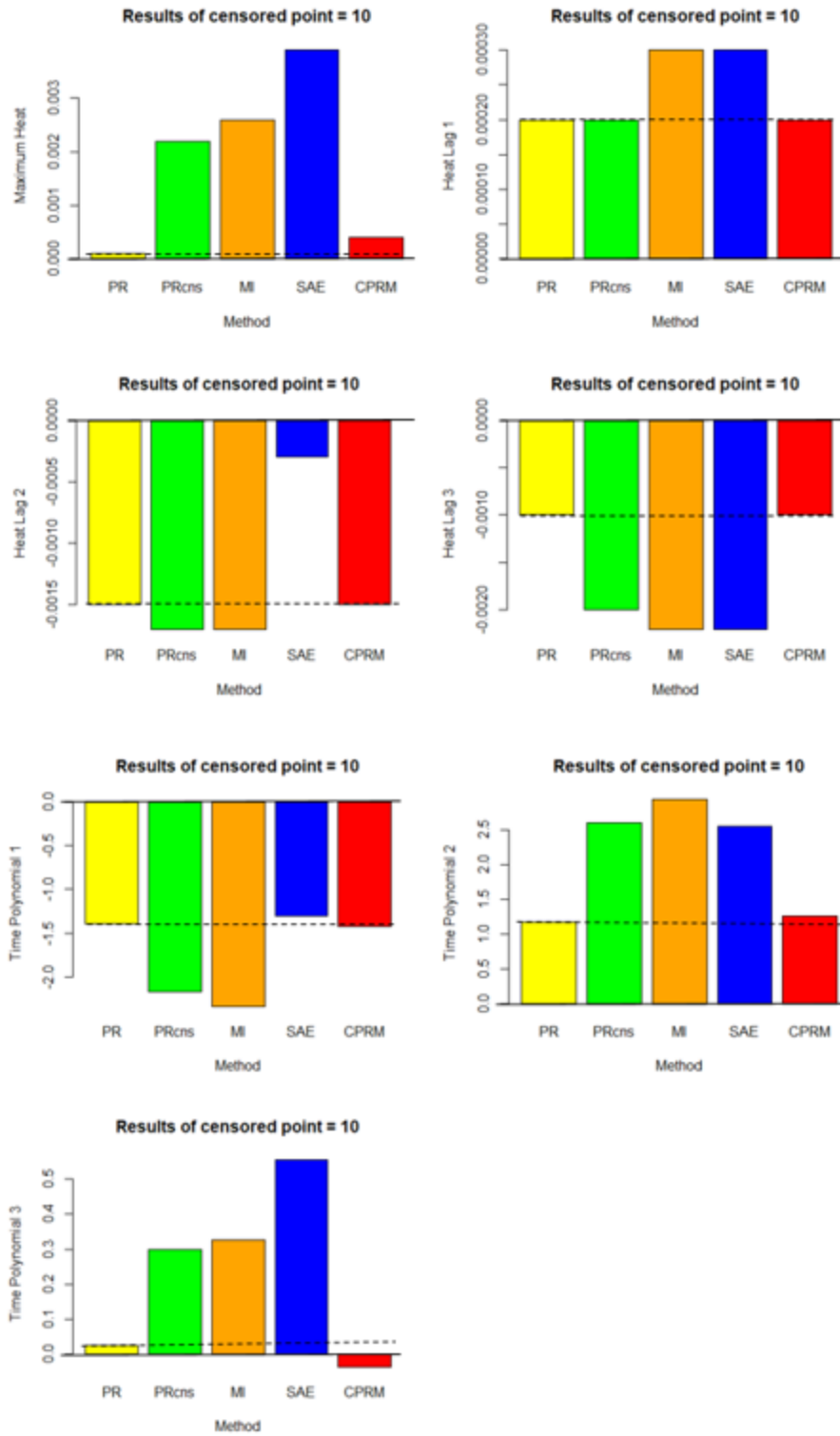


Figure 4. Bar plot for estimations under censored point = 10 (censored proportion = 20.41%)

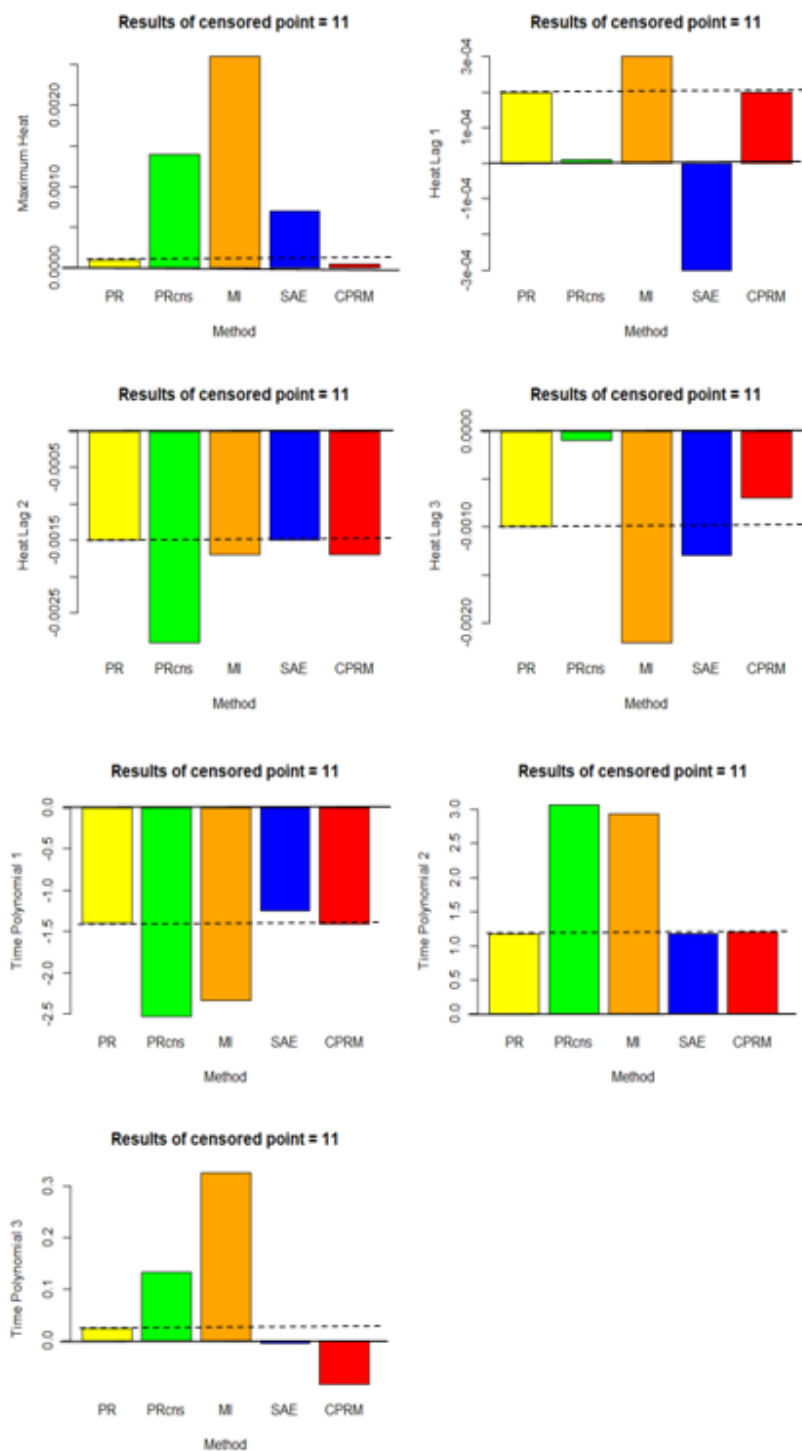


Figure 5. Bar plot for estimations under censored point = 11 (censored proportion = 28.07%)

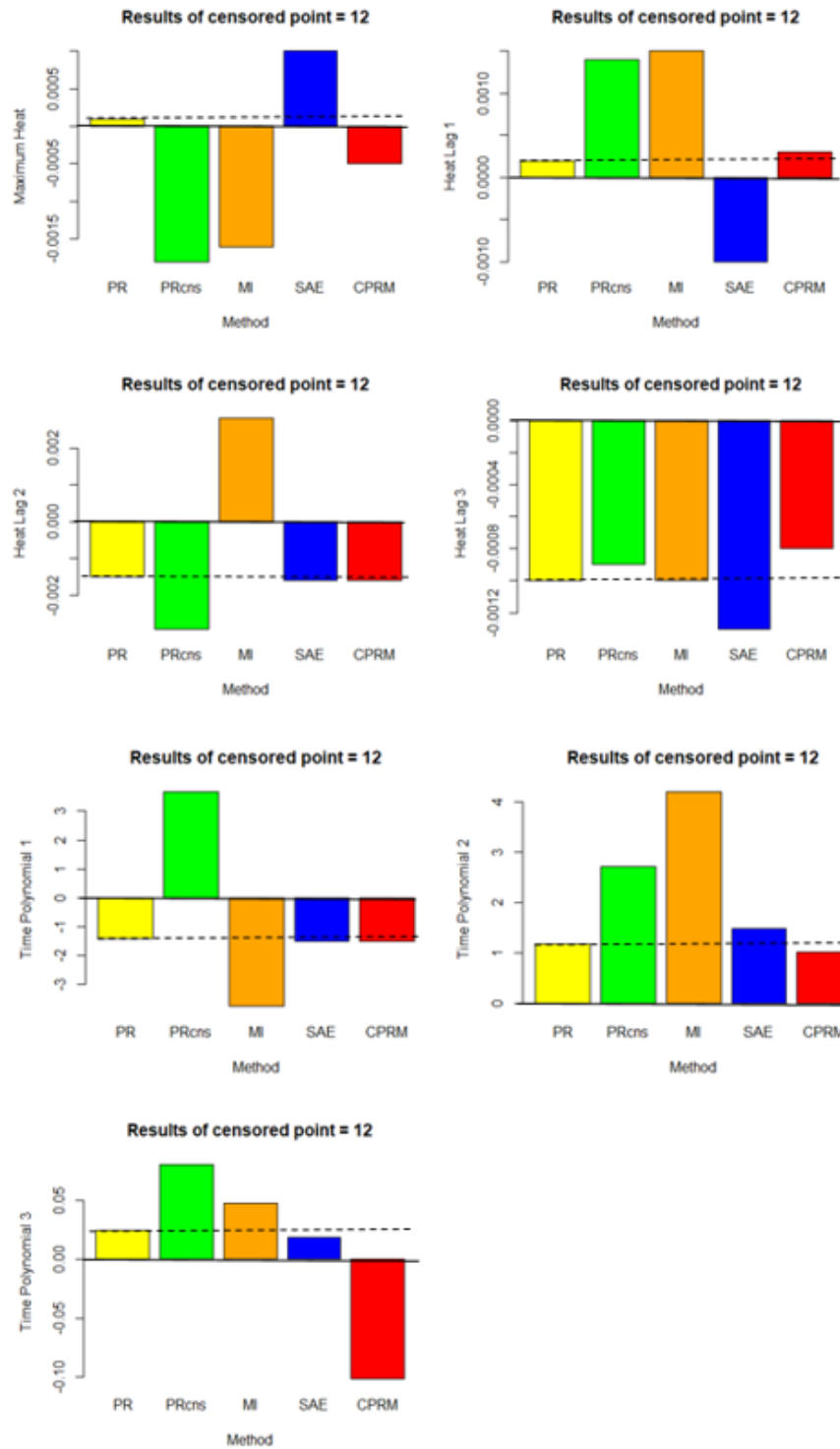


Figure 6. Bar plot for estimations under censored point = 12 (censored proportion = 37.88%)

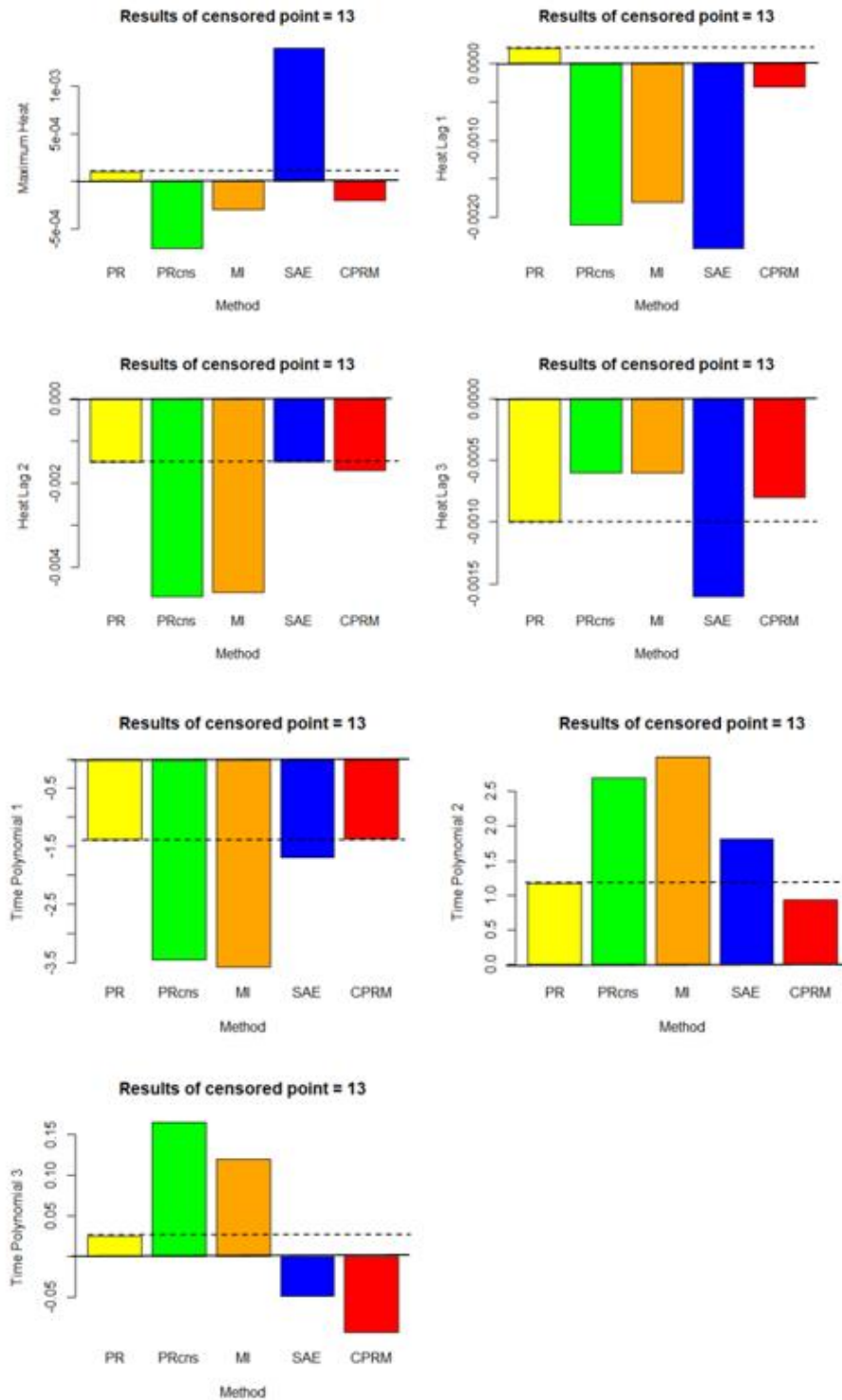


Figure 7. Bar plot for estimations under censored point = 13 (censored proportion = 48.65%)

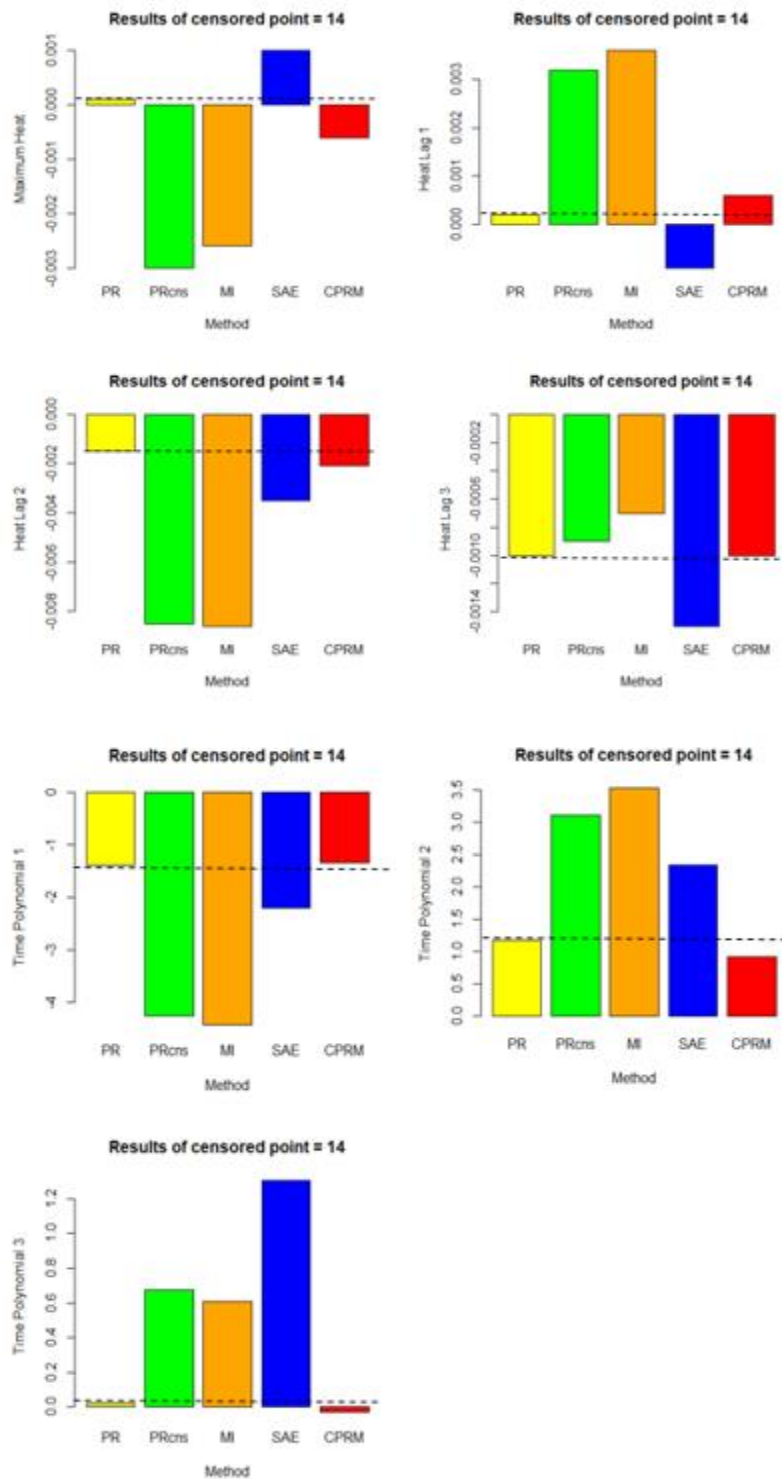


Figure 8. Bar plot for estimations under censored point = 14 (censored proportion = 59.33%)

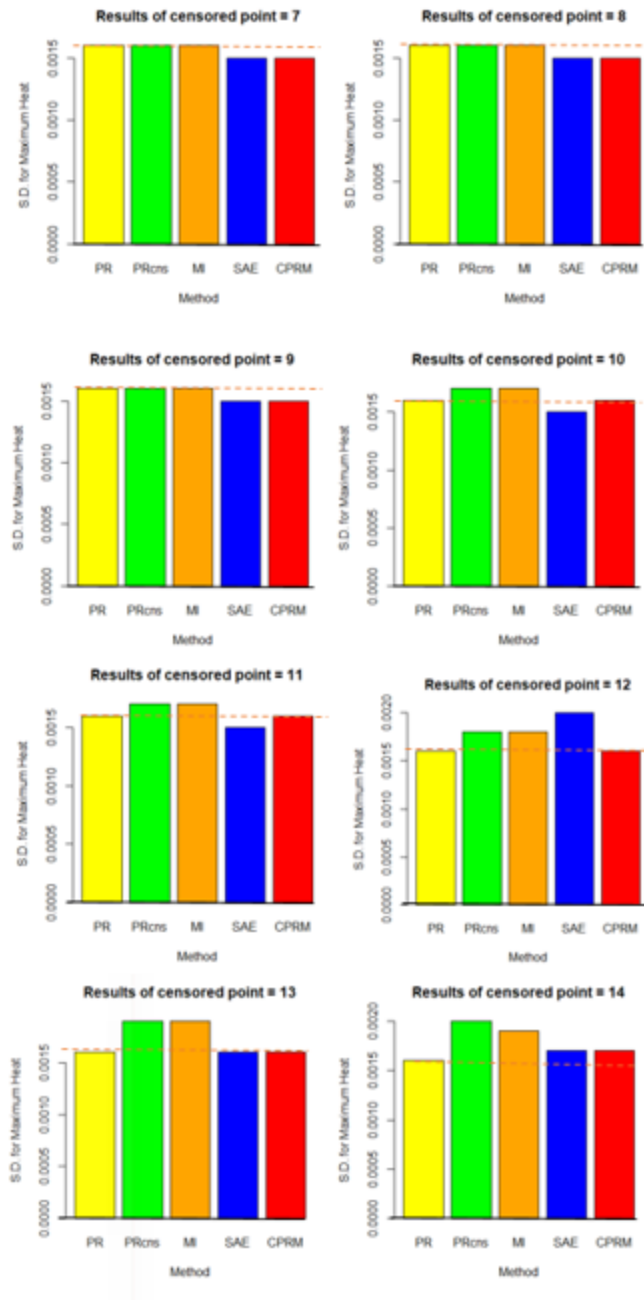


Figure 9. Bar plot for S.D. of Maximum heat estimates under different censored proportions

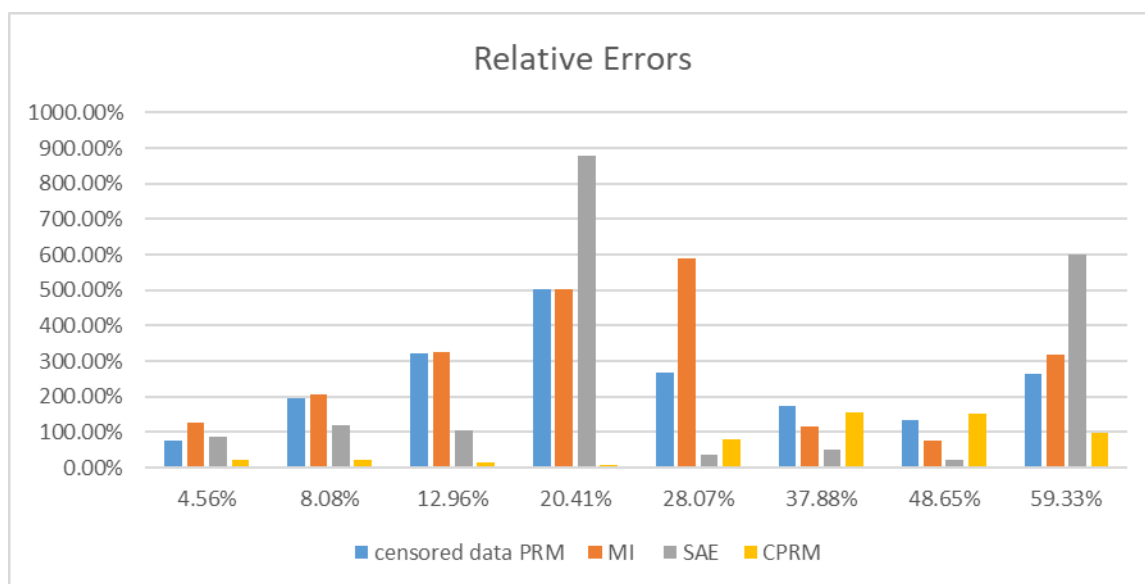


Figure 10. Relative errors under different censored proportion across methods

Appendix

Figures for Journal Article #2

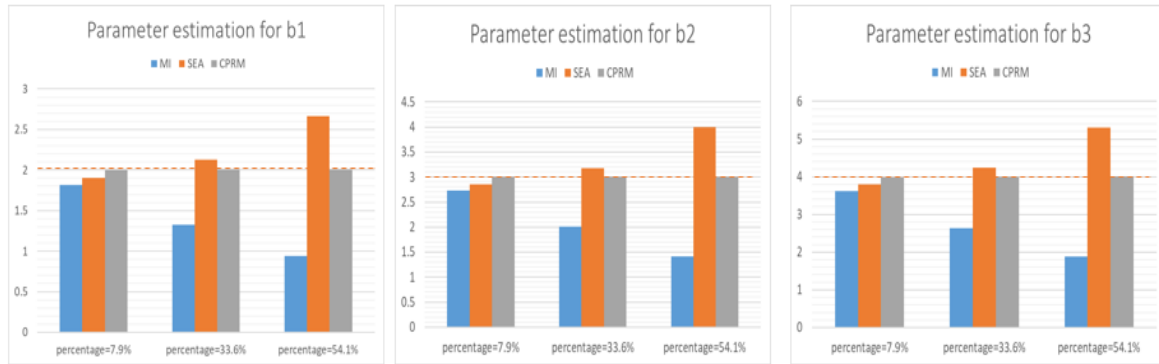


Figure 1. Bar plot for parameter estimation for different methods under different proportions

Note: The red dashed line shows the true value.

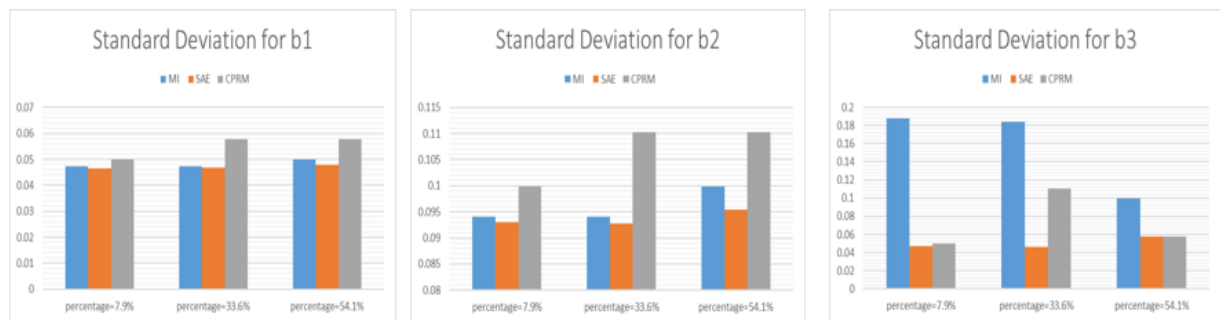


Figure 2. Bar plot for standard deviation for parameter estimation for different methods under different proportions

Appendix

Tables for Journal Article #2

| | Poisson Regression Model | | Censored Data PRM | | MI Method | | Small Area Estimation | | CPRM | |
|---------------------------|--------------------------|--------|-------------------|--------|-----------|--------|-----------------------|--------|---------|--------|
| Censored proportion 4.56% | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | 0.0004 | 0.0016 | 0.0005 | 0.0016 | 0.0001 | 0.0015 | 0.0003 | 0.0015 |
| lag1heat | 0.0002 | 0.0019 | -0.0005 | 0.0019 | 0.0001 | 0.0019 | -0.0001 | 0.0019 | 0.0001 | 0.0019 |
| lag2heat | -0.0015 | 0.0019 | -0.0016 | 0.0019 | -0.0017 | 0.0019 | -0.0010 | 0.0019 | -0.0014 | 0.0019 |
| lag3heat | -0.0010 | 0.0016 | -0.0011 | 0.0016 | -0.0011 | 0.0016 | -0.0008 | 0.0015 | -0.0012 | 0.0015 |
| time1 | -1.3970 | 0.2980 | -1.5830 | 0.3002 | -1.6239 | 0.2986 | -1.2080 | 0.2952 | -1.4238 | 0.2978 |
| time2 | 1.1800 | 0.5821 | 1.3810 | 0.5858 | 1.4838 | 0.5810 | 1.0480 | 0.5771 | 1.2375 | 0.5818 |
| time3 | 0.0246 | 0.2961 | 0.1586 | 0.0983 | 0.1404 | 0.2979 | -0.0684 | 0.2933 | 0.0189 | 0.2951 |

Table 1. Case study results under censored point = 7 (censored proportion = 4.56%)

| | Poisson Regression Model | | Censored Data PRM | | MI Method | | Small Area Estimation | | CPRM | |
|---------------------------|--------------------------|--------|-------------------|--------|-----------|--------|-----------------------|--------|---------|--------|
| Censored proportion 8.08% | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | 0.0006 | 0.0016 | 0.0007 | 0.0016 | 0.0002 | 0.0015 | 0.0003 | 0.0015 |
| lag1 heat | 0.0002 | 0.0019 | -0.0005 | 0.0020 | -0.0004 | 0.0020 | -0.0001 | 0.0019 | 0.0001 | 0.0019 |
| lag2 heat | -0.0015 | 0.0019 | -0.0008 | 0.0020 | -0.0007 | 0.0020 | -0.0012 | 0.0019 | -0.0013 | 0.0019 |
| lag3 heat | -0.0010 | 0.0016 | -0.0018 | 0.0016 | -0.0018 | 0.0016 | -0.0007 | 0.0015 | -0.0012 | 0.0015 |
| time1 | -1.3970 | 0.2980 | -1.7140 | 0.3025 | -1.7571 | 0.3010 | -1.1450 | 0.2983 | -1.4155 | 0.2982 |
| time2 | 1.1800 | 0.5821 | 1.5920 | 0.5898 | 1.7024 | 0.5850 | 0.9717 | 0.5746 | 1.2385 | 0.5825 |
| time3 | 0.0246 | 0.2961 | 0.3017 | 0.0301 | 0.2829 | 0.3001 | -0.1460 | 0.2919 | 0.0208 | 0.2955 |

Table 2. Case study results under censored point = 8 (censored proportion = 8.08%)

| | Poisson Regression Model | | Censored Data PRM | | MI Method | | Small Area Estimation | | CPRM | |
|-----------------------------|--------------------------|--------|-------------------|--------|-----------|--------|-----------------------|--------|---------|--------|
| Censored proportion =12.96% | Mean | S.D. | Mean | S.D. | Mean | S.D | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | 0.0014 | 0.0016 | 0.0015 | 0.0016 | 0.0002 | 0.0015 | 0.0004 | 0.0016 |
| lag1 heat | 0.0002 | 0.0019 | -0.0003 | 0.0020 | -0.0003 | 0.0020 | 0.0000 | 0.0019 | 0.0001 | 0.0019 |
| lag2 heat | -0.0015 | 0.0019 | -0.0014 | 0.0020 | -0.0013 | 0.0020 | -0.0011 | 0.0019 | -0.0014 | 0.0019 |
| lag3 heat | -0.0010 | 0.0016 | -0.0027 | 0.0016 | -0.0027 | 0.0016 | -0.0004 | 0.0015 | -0.0013 | 0.0015 |
| time1 | -1.3970 | 0.2980 | -1.6360 | 0.3060 | -1.6854 | 0.3044 | -1.2150 | 0.2928 | -1.3705 | 0.2989 |
| time2 | 1.1800 | 0.5821 | 1.8320 | 0.5965 | 1.9610 | 0.5912 | 0.9504 | 0.5725 | 1.2258 | 0.5839 |
| time3 | 0.0246 | 0.2961 | 0.2642 | 0.0304 | 0.2452 | 0.3038 | -0.1247 | 0.2909 | -0.0152 | 0.2963 |

Table 3. Case study results under censored point = 9 (censored proportion = 12.96%)

| | Poisson Regression Model | | Censored Data PRM | | MI Method | | Small Area Estimation | | CPRM | |
|-----------------------------|--------------------------|--------|-------------------|--------|-----------|--------|-----------------------|--------|---------|--------|
| Censored proportion =20.41% | Mean | S.D. | Mean | S.D. | Mean | S.D | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | 0.0022 | 0.0017 | 0.0026 | 0.0017 | 0.0039 | 0.0015 | 0.0004 | 0.0016 |
| lag1 heat | 0.0002 | 0.0019 | 0.0002 | 0.0020 | 0.0003 | 0.0021 | 0.0003 | 0.0018 | 0.0002 | 0.0019 |
| lag2 heat | -0.0015 | 0.0019 | -0.0017 | 0.0020 | -0.0017 | 0.0021 | -0.0003 | 0.0018 | -0.0015 | 0.0019 |
| lag3 heat | -0.0010 | 0.0016 | -0.0020 | 0.0016 | -0.0022 | 0.0017 | -0.0022 | 0.0014 | -0.0010 | 0.0016 |
| time1 | -1.3970 | 0.2980 | -2.1660 | 0.3140 | -2.3319 | 0.3149 | -1.3022 | 0.2766 | -1.4219 | 0.3011 |
| time2 | 1.1800 | 0.5821 | 2.6070 | 0.6114 | 2.9389 | 0.6108 | 2.5545 | 0.5382 | 1.2624 | 0.5876 |
| time3 | 0.0246 | 0.2961 | 0.2998 | 0.3125 | 0.3261 | 0.3143 | 0.5551 | 0.2738 | -0.0365 | 0.2986 |

Table 4. Case study results under censored point = 10 (censored proportion = 20.41%)

| | Poisson Regression Model | | Censored Data PRM | | MI Method | | Small Area Estimation | | CPRM | |
|-----------------------------|--------------------------|--------|-------------------|--------|-----------|--------|-----------------------|--------|---------|--------|
| Censored proportion =28.07% | Mean | S.D. | Mean | S.D. | Mean | S.D | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | 0.0014 | 0.0017 | 0.0026 | 0.0017 | 0.0007 | 0.0015 | 0.0000 | 0.0016 |
| lag1 heat | 0.0002 | 0.0019 | 0.0000 | 0.0021 | 0.0003 | 0.0021 | -0.0003 | 0.0019 | 0.0002 | 0.0020 |
| lag2 heat | -0.0015 | 0.0019 | -0.0029 | 0.0021 | -0.0017 | 0.0021 | -0.0015 | 0.0019 | -0.0017 | 0.0019 |
| lag3 heat | -0.0010 | 0.0016 | -0.0001 | 0.0017 | -0.0022 | 0.0017 | -0.0013 | 0.0015 | -0.0007 | 0.0016 |
| time1 | -1.3970 | 0.2980 | -2.5240 | 0.3226 | -2.3319 | 0.3149 | -1.2490 | 0.2937 | -1.4156 | 0.3035 |
| time2 | 1.1800 | 0.5821 | 3.0640 | 0.6281 | 2.9389 | 0.6108 | 1.1810 | 0.5739 | 1.2037 | 0.5920 |
| time3 | 0.0246 | 0.2961 | 0.1333 | 0.3218 | 0.3261 | 0.3143 | -0.0040 | 0.2918 | -0.0834 | 0.3013 |

Table 5. Case study results under censored point = 11 (censored proportion = 28.07%)

| | Poisson Regression Model | | Censored Data PRM | | MI Method | | Small Area Estimation | | CPRM | |
|------------------------------|--------------------------|--------|-------------------|--------|-----------|--------|-----------------------|--------|---------|--------|
| Censored proportion = 37.88% | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | -0.0018 | 0.0018 | -0.0016 | 0.0018 | 0.0010 | 0.0002 | -0.0005 | 0.0016 |
| lag1 heat | 0.0002 | 0.0019 | 0.0014 | 0.0022 | 0.0015 | 0.0022 | -0.0010 | 0.0019 | 0.0003 | 0.0020 |
| lag2 heat | -0.0015 | 0.0019 | -0.0029 | 0.0022 | 0.0028 | 0.0022 | -0.0016 | 0.0019 | -0.0016 | 0.0020 |
| lag3 heat | -0.0010 | 0.0016 | -0.0009 | 0.0018 | -0.0010 | 0.0018 | -0.0013 | 0.0016 | -0.0008 | 0.0016 |
| time1 | -1.3970 | 0.2980 | 3.6630 | 0.3387 | -3.7373 | 0.3370 | -1.5177 | 0.2981 | -1.5095 | 0.3087 |
| time2 | 1.1800 | 0.5821 | 2.7160 | 0.6585 | 4.1919 | 0.0011 | 1.4866 | 0.5822 | 1.0145 | 0.6022 |
| time3 | 0.0246 | 0.2961 | 0.0800 | 0.3390 | 0.0477 | 0.3385 | 0.0180 | 0.2965 | -0.1012 | 0.3077 |

Table 6. Case study Results under censored point = 12 (censored proportion = 37.88%)

| | Poisson Regression Model | | Censored Data PRM | | MI Method | | Small Area Estimation | | CPRM | |
|------------------------------|--------------------------|--------|-------------------|--------|-----------|--------|-----------------------|--------|---------|--------|
| Censored proportion = 48.65% | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | -0.0007 | 0.0019 | -0.0003 | 0.0019 | 0.0014 | 0.0016 | -0.0002 | 0.0016 |
| lag1 heat | 0.0002 | 0.0019 | -0.0021 | 0.0023 | -0.0018 | 0.0023 | -0.0024 | 0.0020 | -0.0003 | 0.0020 |
| lag2 heat | -0.0015 | 0.0019 | -0.0047 | 0.0023 | -0.0046 | 0.0023 | -0.0015 | 0.0020 | -0.0017 | 0.0020 |
| lag3 heat | -0.0010 | 0.0016 | -0.0006 | 0.0019 | -0.0006 | 0.0019 | -0.0016 | 0.0016 | -0.0008 | 0.0016 |
| time1 | -1.3970 | 0.2980 | -3.4620 | 0.3592 | -3.5757 | 0.3572 | -1.7025 | 0.3068 | -1.3711 | 0.3168 |
| time2 | 1.1800 | 0.5821 | 2.6910 | 0.6976 | 2.9978 | 0.6907 | 1.8194 | 0.5990 | 0.9357 | 0.6170 |
| time3 | 0.0246 | 0.2961 | 0.1648 | 0.3598 | 0.1190 | 0.3592 | -0.0493 | 0.3057 | -0.0933 | 0.3152 |

Table 7. Case study results under censored point = 13 (censored proportion = 48.65%)

| | Poisson Regression Model | | Censored Data PRM | | MI Method | | Small Area Estimation | | CPRM | |
|------------------------------|--------------------------|--------|-------------------|--------|-----------|--------|-----------------------|--------|---------|--------|
| Censored proportion = 59.33% | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Max Heat | 0.0001 | 0.0016 | -0.0030 | 0.0020 | -0.0026 | 0.0019 | 0.0010 | 0.0017 | -0.0006 | 0.0017 |
| lag1 heat | 0.0002 | 0.0019 | 0.0032 | 0.0025 | 0.0036 | 0.0023 | -0.0009 | 0.0021 | 0.0006 | 0.0021 |
| lag2 heat | -0.0015 | 0.0019 | -0.0085 | 0.0025 | -0.0086 | 0.0023 | -0.0035 | 0.0021 | -0.0021 | 0.0021 |
| lag3 heat | -0.0010 | 0.0016 | -0.0009 | 0.0020 | -0.0007 | 0.0019 | -0.0015 | 0.0017 | -0.0010 | 0.0017 |
| time1 | -1.3970 | 0.2980 | -4.2540 | 0.3896 | -4.4082 | 0.3572 | -2.1990 | 0.3224 | -1.3410 | 0.3294 |
| time2 | 1.1800 | 0.5821 | 3.1030 | 0.7530 | 3.5177 | 0.6907 | 2.3290 | 0.6281 | 0.9130 | 0.6398 |
| time3 | 0.0246 | 0.2961 | 0.6726 | 0.3901 | 0.6085 | 0.3592 | 1.3030 | 0.3216 | -0.0260 | 0.3276 |

Table 8. Case study results under censored point = 14 (censored proportion = 59.33%)

References

- [1] Cordero A, Mulinare J, Berry R, Boyle C, Dietz W, Johnston Jr R, et al. CDC Grand Rounds: additional opportunities to prevent neural tube defects with folic acid fortification. *Morbidity and mortality weekly report*. 2010;59:980-4.
- [2] Prevention CfDCa. Lyme Disease Home. 2017.
- [3] Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. *Journal of statistical software*. 2008;27:1-25.
- [4] Rosamond W, Flegal K, Furie K, Go A, Greenlund K, Haase N, et al. Heart disease and stroke statistics—2008 update. *Circulation*. 2008;117:e25-e146.
- [5] Whitehead NS, Leiker R. Case management protocol and declining blood lead concentrations among children. *Prev Chronic Dis* [serial online]. 2007.
- [6] Dignam TA, Lojo J, Meyer PA. Reduction of elevated blood lead levels in children in North Carolina and Vermont, 1996–1999. *Environmental health perspectives*. 2008;116:981.
- [7] Schmee J, Hahn GJ. A simple method for regression analysis with censored data. *Technometrics*. 1979;21:417-32.
- [8] Shih WJ. Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine*. 2002;3:4.
- [9] Jin X, Carlin BP, Banerjee S. Generalized hierarchical multivariate CAR models for areal data. *Biometrics*. 2005;61:950-61.
- [10] Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581-92.
- [11] Rubin DB. *Multiple imputation for nonresponse in surveys*: John Wiley & Sons; 2004.
- [12] Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. *Statistics in medicine*. 2001;20:1541-9.

- [13] Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research*. 2015;24:462-87.
- [14] Ymann C, Wrbach A, Gomann S, Geissler F, Bela A. Nonparametric Multiple Imputation for Questionnaires with Individual Skip Patterns and Constraints: The Case of Income Imputation in the National Educational Panel Study. *Sociological Methods & Research*. 2017;46:864-97.
- [15] Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data. *Jama*. 2015;314:1966-7.
- [16] Eekhout I, de Vet HC, Twisk JW, Brand JP, de Boer MR, Heymans MW. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of clinical epidemiology*. 2014;67:335-42.
- [17] Enders CK, Mistler SA, Keller BT. Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological methods*. 2016;21:222.
- [18] Cadwell BL, Thompson TJ, Boyle JP, Barker LE. Bayesian small area estimates of diabetes prevalence by US county, 2005. *Journal of Data Science*. 2010;8:171-88.
- [19] Barker LE, Thompson TJ, Kirtland KA, Boyle JP, Geiss LS, McCauley MM, et al. Bayesian small area estimates of diabetes incidence by United States county, 2009. *Journal of data science: JDS*. 2013;11:269.
- [20] Zhang X, Holt JB, Lu H, Wheaton AG, Ford ES, Greenlund KJ, et al. Multilevel regression and poststratification for small-area estimation of population health outcomes: a

case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American journal of epidemiology*. 2014;179:1025-33.

[21] Dwyer-Lindgren L, Flaxman AD, Ng M, Hansen GM, Murray CJ, Mokdad AH.

Drinking patterns in US counties from 2002 to 2012. *American journal of public health*. 2015;105:1120-7.

[22] Terza JV. A Tobit-type estimator for the censored Poisson regression model. *Economics Letters*. 1985;18:361-5.

[23] Famoye F, Wang W. Censored generalized Poisson regression model. *Computational statistics & data analysis*. 2004;46:547-60.

[24] Mahmoud MM, Alderiny MM. On estimating parameters of censored generalized Poisson regression model. *Applied Mathematical Sciences*. 2010;4:623-35.

[25] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*. 2009;71:319-92.