2019

A Flexible Zero-Inflated Poisson Regression Model

Eric S. Roemmele
University of Kentucky, eric.roemmele@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2019.187

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Eric S. Roemmele, Student

Dr. Derek Young, Major Professor

Dr. Constance Wood, Director of Graduate Studies

A Flexible Zero-Inflated Poisson Regression Model

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Eric Roemmele
Lexington, Kentucky

Co-Directors: Dr. Derek Young, Professor of Statistics
and Dr. Richard Kryscio, Professor of Statistics
Lexington, Kentucky 2019

ABSTRACT OF DISSERTATION

A Flexible Zero-Inflated Poisson Regression Model

A practical problem often encountered with observed count data is the presence of excess zeros. Zero-inflation in count data can easily be handled by zero-inflated models, which is a two-component mixture of a point mass at zero and a discrete distribution for the count data. In the presence of predictors, zero-inflated Poisson (ZIP) regression models are, perhaps, the most commonly used. However, the fully parametric ZIP regression model could sometimes be restrictive, especially with respect to the mixing proportions. Taking inspiration from some of the recent literature on semiparametric mixtures of regressions models for flexible mixture modeling, we propose a semiparametric ZIP regression model. We present an "EM-like" algorithm for estimation and a summary of asymptotic properties of the estimators. The proposed semiparametric models are then applied to a data set involving clandestine methamphetamine laboratories and Alzheimer's disease.

KEYWORDS: Bootstrap; Count data; EM Algorithm; zero-inflation; semiparametric model

Author's signature: Eric Roemmele

Date: May 6, 2019

A Flexible Zero-Inflated Poisson Regression Model

By

Eric Roemmele

Co-Director of Dissertation: Derek Young

Co-Director of Dissertation: Richard Kryscio

Director of Graduate Studies: Constance Wood

Date: May 6, 2019

Dedicated to my family

ACKNOWLEDGMENTS

I would like to start by thanking my advisor, Dr. Derek Young, for his guidance on this research. Furthermore, thank you to my other committee members Dr. David Allen, Dr. Richard Kryscio, Dr. Richard Charnigo, and Dr. Carlos Lamarche for their service and suggestions. Moreover, thanks to the Univeristy of Kentucky Statistics Department for providing me with a highly enjoyable five years, and in particular, Dr. Arnold Stromberg for his insights and support. Lastly, the University of Kentucky Department of Agriculture, Food, and the Environment has allowed me to work with terrific investigators in a plethora of fields, in addition to supporting me financially.

Next, without the love and support of my family, this work would not be possible. I have been blessed with such a wonderful family. My mother has worked extremely hard throughout her whole life to support three children, along with fostering our moral and educational needs. Dr. Melissa Roemmele, my sister, has been a great role model, demonstrating compassion and guidance when I was struggling. My brother, Brian, always makes me laugh, and is one of my best friends. Finally, my stepfather, Frank Hsu, has taught me a great deal about a variety of life issues.

Finally, the University of Dayton Mathematics Department always challenged me to be a better student and mathematician, and prepared me well for graduate school. In particular, thank you to Dr. Lynne Yengulalp and Dr. Joseph Mashburn for challanging, but fruitful coursework. Thank you to my friends Kelsi Bertrams, Sam Bertrams, Ryan Kelly, Matt McNeil, Tracy Moor, Kyle McGrail, and Thomas Feighery for being great friends. Moreover, thank you to Matt Rutledge, Jennifer Daddysmen, Woody Burchett, Kristen McQuery, Aviv Brokman, Josh Lambert, and Sarah Janse for making graduate school such an enjoyable experience. I would like to

finish by thanking Paul Walker, Dwayne Johnson, Ryan Gosling, Jon Hamm, Bono, Bruce Springsteen, Fyodor Dostoevsky and Bob Dylan.

TABLE OF CONTENTS

Acknowledgments iii

List of Tables vii

List of Figures viii

Chapter 1 Introduction to Zero-Inflated Models 1
1.1 Introduction 1
1.2 Traditional ZI Models 3
1.3 Estimation and Inference 4
1.4 Bayesian ZI Models 11
1.5 Software and Numerical Demonstrations 20
1.6 Example: Insurance Data 22
1.7 ZI Count Regression Models for Handling Data Dispersion 26
1.8 ZI Models for Clustered Data 28
1.9 Zero-Inflation and Diagonal-Inflation in Multivariate Count Responses 35
1.10 Related Models 35
1.11 Example: Relationship Data 38
1.12 Appendix 41

Chapter 2 Semiparametric Extension to ZIP Regression via Local Likelihood 48
2.1 Introduction 48
2.2 Estimation of Semiparametric Regression Models 52
2.3 Inference 64
2.4 Simulation Studies 69
2.5 Real Data Analysis 74
2.6 Appendix 82

Chapter 3 Conclusions and Future Directions 99
3.1 Conclusions 99
3.2 Future Directions 100
3.3 Zero-Inflation in Spatial Data 102

Bibliography 106

Vita 121

LIST OF TABLES

1.1 ZIP Software Estimation Comparison 20
1.2 ZINB Software Estimation Comparison 21
1.3 Bins of Vehicle Body Types 23
1.4 Count Component Variables Selected 24
1.5 BIC and AIC Values for Fitted Models 24
1.6 ZIP Model Coefficients 26
1.7 Adjusted-R^2 Table 40
1.8 AIC and BIC Values for Couple Data 41
1.9 ZIP Timing Results 44
1.10 ZINB Timing Results 44

2.1 Common Kernel Functions 55
2.2 Accuracy of Estimates Table 70
2.3 Beta Coverage 72
2.4 Coverage rates of intervals for π 73
2.5 Inclusion Model Comparison 76
2.6 Regression Coefficients for Meth Data 79
2.7 Accuracy of Estimates for 2^{nd} Simulation 95
2.8 Beta Coverage for Simulation Study 2 96
2.9 Coverage rates of intervals for π 97

3.1 Spatial Estimates 104
3.2 Prediction accuracy on 2012 data set. 104
3.3 Predictions for Fayette and Jefferson County in 2012. 105

LIST OF FIGURES

1.1 Equadispersion compared to Overdispersion 5
1.2 Histogram of the Number of Claims 22
1.3 Randomized Quantile Residual Plots 25
1.4 Couple Data Histogram 39
1.5 Q–Q Plots for Couple Data Fits 39

2.1 Estimated Curve for $\pi(x)$ 70
2.2 Accuracy of Estimates Plot 71
2.3 BC Intervals for π when $n = 400$ 74
2.4 Histogram of bootstrap LRT statistics for a single Monte-Carlo replicate 75
2.5 Simulated Power Functions of the LRT 76
2.6 Histogram of Subiculum Inclusions. Histogram is truncated at five. . . . 77
2.7 Partial Dependency Plots for Alzheimer's Data 78
2.8 Zero-inflation Probability against MMSE 78
2.9 Histogram of lab seizures truncated at 20. 80
2.10 Heat Map for KY 80
2.11 Partial dependence of lab counts on median earnings 81
2.12 Estimated Zero-Inflation by PSE Sales 81
2.13 Estimated Zero-Inflation for Simulation Study 2 98

Chapter 1 Introduction to Zero-Inflated Models

1.1 Introduction

Poisson and negative binomial (NB) regression models are two of the most common models for count data. However, the behavior of the zero counts in the observed data may create difficulties for these models. For example, zero counts may be impossible, which is known as *zero-truncation* [1]. Or, the zero and non-zero counts may be generated through differerent processes, for which *hurdle models* are commonly employed [2]. Moreover, observed data may exhibit *zero-inflation*, which is when the observed data have excessive zeros relative to the assumed count distribution. Mathematically, zero-inflated (ZI) models are two-component mixtures of a point mass at zero and a count distribution. The seminal paper by [3] is among the earliest works to develop ZI regression models in the prescence of covariates. In that paper, the ZI Poisson (ZIP) regression model is introduced and applied to model the number of defects in a soldered switchboard. The two components (degenerate and count) are interpreted as a *perfect* state where defects are impossible, and an *imperfect* state where defects are possible, respectively.

ZI regression models are heavily utilized across various disciplines, including ecology. For example, [4] analyzed avian abundance by ZI regression models, while discussing how zeros arise in an ecological dataset. In detail, the authors define *true zero counts* and *false zero counts*, which are the zeros from the degenerate and count component, respectively. A true zero count may arise because the species does not occur at the site, while a false zero count may arise because the species occurs at a site, but is not present during the survey period or the observer failed to detect it. Other ecological applications of ZI regression in fish abundance and vulnerable plant species abundance can be seen in [5] and [6], respectively. The general importance of ZI regression models in ecology is also emphasized in Chapter 11 of [7].

Another discipline where ZI regression models are frequently employed is insurance. [8] paralleled the notion of *strategic* and *incidental* zeros to the perfect and imperfect states introduced in [3]. For instance, in determining the health policy of a (potential) policyholder, the number of physician visits can be an indication of overall health and, thus, affect the level of policy coverage. [8] noted that when modeling the number of physician visits, a person might have zero visits during a time period (or *exposure*) because they follow alternative medicine and never visit a physician

(strategic zero), or because they were healthy during the time period and had no reason to visit a physician (incidental zero). [9] noted similar states with how zeros in the no claim discount (NCD) system, which is widely used by automobile insurers. Policyholders could have zero claims either because they typically will not file the claim if it doesn't meet the deductible (strategic zero), or because they simply did not have any issues regarding their automobile (incidental zero). Similar analyses involving ZI regression modeling for risk classification of claim counts are performed in [10, 11, 12].

While ecology and insurance are two fields where ZI regression models are commonly used, the utility of such models has been demonstrated with numerous other diverse applications. Examples highlighting the broad range of interesting applications using ZI regression models include the development of a functional relationship between truck accidents and the geometric design of road sections [13], an analysis of economists seeking academic interviews after tenure denial [14], a study about the effects of cigarette price change on smoking behavior [15], the assessment of dental cavities in low birth weight adolescents [16], and development of a set of models to characterize the on-going quality of census frames in preparation for the 2020 United States Census [17].

The rest of this chapter is organized as follows: Section 1.2 defines the zero-inflated mass function along with the zero-inflated regression model. In the framework of [3], Section 1.3 provides a formal discussion of the ZI count regression model with emphasis on the two most commonly assumed discrete distributions for the count component - the Poisson and negative binomial distributions. In this section, estimation and inference is discussed from a frequentist perspective in greater detail. Section 1.4 discusses Bayesian ZI regression models and Bayesian diagnostic developments. Section 1.5 reviews software in `SAS` and `R` for fitting ZI regression models. Section 1.6 provides an example of several count regression and ZI regression fits on an auto insurance data set, along with model comparisons and diagnostics. Section 1.7 discusses zero-inflation in the context of modeling with non-standard discrete distributions when data dispersion is present. Section 1.8 highlights developments of analyzing correlated (or clustered) ZI counts, such as longitudinal data and time series data. Section 1.9 briefly discusses multivariate ZI regression models and their applications. Section 1.10 discusses models that are related to ZI regression models, such as zero-truncated and hurdle models. Section 1.11 examines different ZI regression models applied to a unwanted pursuit behavior data set. Finally, Section 1.12 is an appendix where the ECM algorithm for ZI negative binomial regression

is presented, as well as `JAGS` and `R` code for estimation of a Bayesian ZIP regression model.

1.2 Traditional ZI Models

Let the discrete random variable $Y \in \mathbb{N}$ be a count of interest; e.g., length of stay in the hospital [18], the counts of trees in a forest using grid-cell data [19], or the number of added or deleted housing units in a census block [17]. Moreover, let $\mathcal{C}$ be a latent class indicator such that the conditional distribution

$$Y|\mathcal{C} = c \sim \begin{cases} p(y; \mu, \boldsymbol{\vartheta}) & c = 0 \\ 0 & c = 1 \end{cases}.$$

Here, $p(y; \mu, \boldsymbol{\vartheta})$ is a mass function on $\mathbb{N}$ with mean μ, and $\boldsymbol{\vartheta}$ pertains to any additional (possibly nuissance) parameters, such as the heterogeneity parameter in negative binomial regression. Now, $\mathcal{C}$ is unobservable, and so we seek to model the marginal distribution of Y. The marginal distribution of Y is

$$\begin{aligned} f_Y(y; \mu, \boldsymbol{\vartheta}) &= \mathbb{P}(\mathcal{C} = 1)\mathbb{P}(Y = y|\mathcal{C} = 1) + \mathbb{P}(\mathcal{C} = 0)\mathbb{P}(Y = y|\mathcal{C} = 0) \\ &= \pi \ \mathrm{I}\{y = 0\} + (1 - \pi)p(y; \mu, \boldsymbol{\vartheta}), \end{aligned} \tag{1.1}$$

where $\pi = \mathbb{P}(\mathcal{C} = 1)$. The quantity π is called the mixing proportion, latent class probability, or more commonly, the zero-inflation probability. The mass function in Equation 1.1 is called the ZI mass function with count distribution $p(\cdot)$, which is a mixture of a degerate distribution at zero and a count distribution. So, if there is a positive count, then that observation must come from $p(\cdot)$. But, a observed zero could come from the degenerative state or a random zero observed from $p(\cdot)$. In interpretation, suppose Y is the amount of times a individual visited a doctor in a calender year. If a patient had no medical issues in a calender year, that zero is more likely to be from the degenerative state since that patient is not at risk. On the other hand, a patient could have zero doctor visits even though they were at risk, i.e. had health ailments, but didn't want to pay the bill to go to the doctor, or they were on an alternative medicine plan. This observed zero is more likely to a random zero from the count component $p(\cdot)$.

The extension to ZI regression is analagous to that of *generalized linear models* (GLM). Suppose that $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{W} \in \mathbb{R}^q$ are vectors of covariates measured with the response Y. Let $\boldsymbol{x}$, $\boldsymbol{w}$, and y be, respectively, the realizations of those

variables. Suppose the mean function for the count component can be related to $\boldsymbol{x}$ by $g(\mu) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$ for some link function $g(\cdot)$, and the zero-inflation probabilities are related to covariates through $h(\pi) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha}$ via the link function $h(\cdot)$. Typically, $g(\cdot)$ is taken as the log link, and $h(\cdot)$ is taken to be the logit link. Also, the count of interest is sometimes measured in terms of its exposure, N. Assuming a log-link implies

$$\log(\mu) = \log(N) + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}, \tag{1.2}$$

where $\log(N)$ appears as an offset term in the right-hand side of the above expression. For the remainder of this chapter, we'll write $\pi(\boldsymbol{\alpha}) = h^{-1}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha})$ and $\mu(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$ to denote the zero-inflation probabilities and mean of the count distribution, respectively.

The ZI Regression model then can be written as the two-component mixture model

$$Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w} \sim \pi(\boldsymbol{\alpha})\mathrm{I}\{y = 0\} + (1 - \pi(\boldsymbol{\alpha}))p(y; \mu(\boldsymbol{\beta}), \boldsymbol{\vartheta}). \tag{1.3}$$

Note that the predictors $\boldsymbol{w}$ may be uncoupled from those predictors in $\boldsymbol{x}$. Lastly, it is typical to assume that $\mu(\boldsymbol{\beta})$ and $\pi(\boldsymbol{\alpha})$ are not functionally related, although this need not be the case [3].

1.3 Estimation and Inference

The most common count distributions used in ZI regression models [3, 20, 21] are the Poisson, which has pmf

$$p(y; \mu) = (y!)^{-1}\mu^y e^{-\mu} \ \mu > 0, \tag{1.4}$$

and the negative binomial, which employing the gamma-Poisson (mixture) representation, has pmf

$$p(y; \mu, \theta) = \frac{\Gamma(\theta + y)}{y!\Gamma(\theta)}\Big(\frac{\mu}{\theta + \mu}\Big)^y\Big(\frac{\theta}{\theta + \mu}\Big)^\theta \ \ (\mu, \theta > 0), \tag{1.5}$$

where μ is the mean and θ is called the *heterogeneity* or *dispersion* parameter. Here, $\Gamma(\cdot)$ denotes the Gamma function

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-t}dx.$$

The Poisson model assumes *equidispersion*, i.e. the mean and variance are equal. This strong assumption however, is often violated in practice. On the other hand, it can be shown that the negative binomial (NB) distribution has mean $\mathbb{E}(Y) = \mu$ and variance $\text{Var}(Y) = \mu + \mu^2\theta$. Thus, the variance increases quadratically with the mean, and so the negative binomial distribution is often used to model data that exhibit *overdispersion*; i.e. as the mean of the count distribution increases, the variance increases at a faster rate than the mean. A visual representation of Poisson equdispersion and NB overdispersion can be seen in Figure 1.1.

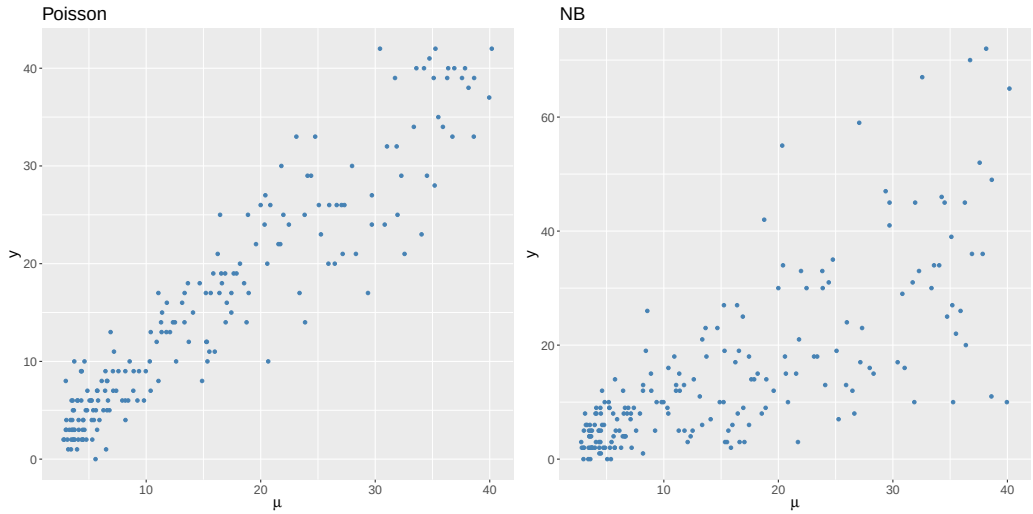


Figure 1.1: Equadispersion (left) versus Overdispersion (right)

In the negative binomial pmf, the heterogeneity parameter θ is usually assummed constant, but could be modeled as a function of covariates if there is overdispersion or underdispersion relative to the negative binomial regression model; see Chapter 7.5 of [22].

Suppose we have a sample size of n. The pmfs in (1.3) and (1.4) can be extended to the Poisson regression pmf and negative binomial pmf by modeling $\log(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, or equivalently, $\mu_i(\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$. We'll assume the mixing proportions can be modeled as $\text{logit}(\pi_i) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha}$, or equivalently, $\pi_i(\boldsymbol{\alpha}) = \text{logit}^{-1}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha})$. Then, using the ZI count regression pmf in (1.3), it follows that the ZIP regression log-likelihood is

$$\begin{aligned}\ell_1(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{y}) = &\sum_{y_i=0} \log\Big(\pi_i(\boldsymbol{\alpha}) + (1-\pi_i(\boldsymbol{\alpha}))\exp\{\mu_i(\boldsymbol{\beta})\}\Big) \\ &+ \sum_{y_i>0}\Big[\log(1-\pi_i(\boldsymbol{\alpha})) - \mu_i(\boldsymbol{\beta}) + y_i\log(\mu_i(\boldsymbol{\beta})) - \log(y!)\Big],\end{aligned} \tag{1.6}$$

and the ZINB regression log-likelihood is

$$\begin{aligned}\ell_2(\boldsymbol{\beta}, \boldsymbol{\alpha}, \theta; \boldsymbol{y}) = &\sum_{y_i=0} \log\left(\pi_i(\boldsymbol{\alpha}) + (1-\pi_i(\boldsymbol{\alpha}))\Big(\frac{\theta}{\theta+\mu_i(\boldsymbol{\beta})}\Big)^{\theta}\right) \\ &+ \sum_{y_i>0}\Big[\log(1-\pi_i(\boldsymbol{\alpha})) + \log(\Gamma(\theta+y_i)) - \log(\Gamma(\theta)) - \log(y_i!) \\ &+ y_i\log(\mu_i(\boldsymbol{\beta})) + \theta\log(\theta) - (\theta+y_i)\log(\theta+\mu_i(\boldsymbol{\beta}))\Big],\end{aligned} \tag{1.7}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. Note that the ZI Geometric (ZIG) regression model is a special case of the ZINB regression model when $\theta = 1$, which has been discussed in detail by [23]. Also, if the response has an upper bounded count, then a ZI binomial (ZIB) regression model is easily obtained; see [24].

Maximum likelihood estimation for ZIP and ZINB regression models has been thoroughly treated in [3] and [22], respectively. Under typical regularity conditions, the asymptotic distribution of the maximum likelihood estimates (MLEs) is, of course, multivariate normal. While closed forms solutions for the MLEs do not exist, they are easily obtained using numerical methods. Moreover, [3] and [25] provide the formulas for the gradients and second derivatives, which can then be used for computing the observed information matrix for standard errors.

EM Algorithms

Newton-Raphson algorithms are one of the most frequently utilized method for calculating the ZIP and ZINB regression MLEs. The Newton-Raphson algorithm, when it converges, is typically faster than the *Expectation-Maximization* (EM) algorithm. However, EM algorithms are quite easy to code and take advantage of the mixture structure of ZI regression models by iteratively fitting weighted versions of simpler GLMs [26].

EM algorithms seek MLEs of parameters in statistical models where the model depends on latent (or unobserved) random variables [27]. In the case of ZI regression, the latent variable is the degerate or count component class membership for the i$^{\text{th}}$ observation. A summary of the EM Algorithm is as follows : Let $\boldsymbol{X}$ be the observed data, and $\boldsymbol{R}$ be the set of missing or latent variables. Let the complete data $(\boldsymbol{X}, \boldsymbol{R}) \sim f(\boldsymbol{X}, \boldsymbol{R}; \boldsymbol{\theta})$, where f is a density (or mass function) and $\boldsymbol{\theta}$ is a vector of parameters. Denote the complete data likelihood function as $L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{R}) = f(\boldsymbol{X}, \boldsymbol{R}; \boldsymbol{\theta})$.

We seek the MLEs of the observed data likelihood

$$L(\boldsymbol{\theta}; \boldsymbol{X}) = f(\boldsymbol{X}; \boldsymbol{\theta}) = \int f(\boldsymbol{X}, \boldsymbol{R}; \boldsymbol{\theta}) d\boldsymbol{R} \tag{1.8}$$

The above likelihood function can be difficult to optimize. Instead, the complete data likelihood, which is typically easier to optimize than the observed likelihood, is utilized as a surrogate. The EM algorithm simiplifies the optimization by iteratively applying E-Steps and M-Steps.

EM Algorithm - For $t = 0, 1, \ldots,$ do:

1. *Expectation Step* (E-Step) : Compute the expectation

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{R}|\boldsymbol{X};\boldsymbol{\theta}^{(t)}} \Big[\log L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{R}) \Big], \tag{1.9}$$

   where the expectation is with respect to the conditonal distribution of $\boldsymbol{R}|\boldsymbol{X}$ and the current estimate of the paramter $\boldsymbol{\theta}^{(t)}$.

2. *Maximization Step* (M-Step) : Maximize (1.9) :

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}). \tag{1.10}$$

Repeat steps 1 and 2 until the change in the marginal likelihood is small (i.e. $L(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{X}) - L(\boldsymbol{\theta}^{(t)}; \boldsymbol{X}) < \epsilon$, where ϵ is sufficiently small).

For the EM algorithm applied to ZIP Regression, let $R_i = \mathrm{I}\{Y_i \text{ from the degenerate state}\}$, and $\boldsymbol{R} = (R_1, \ldots, R_n)^{\mathrm{T}}$.
Therefore, $R_i \sim \text{Bern}(\pi_i(\boldsymbol{\alpha}))$, and

$$Y_i | R_i = r_i \sim \begin{cases} 0 & r_i = 1 \\ \text{Poisson}(\mu_i(\boldsymbol{\beta})) & r_i = 0 \end{cases}. \tag{1.11}$$

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$. Then, the complete data log-likelihood is

$$\begin{aligned}
\ell_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{r}) &= \sum_{i=1}^{n} \log(f_{R_i}(r_i) \times f_{Y_i|R_i}(y_i|r_i)) \\
&= \sum_{i=1}^{n} \log(f_{R_i}(r_i)) + \sum_{i=1}^{n} \log(f_{Y_i|R_i}(y_i|r_i)) \\
&= \sum_{i=1}^{n} \Big[r_i \text{logit}(\pi_i) + \log(1-\pi_i) \Big] + \sum_{i=1}^{n} \text{I}\{r_i = 0\} \Big[y_i \log(\mu_i) - \mu_i - \log(y_i!) \Big] \\
&\propto \sum_{i=1}^{n} \Big[r_i(\boldsymbol{w}_i^{\text{T}} \boldsymbol{\alpha}) - \log(1 + \exp(\boldsymbol{w}_i^{\text{T}} \boldsymbol{\alpha})) \Big] + \sum_{i=1}^{n} (1 - r_i) \Big[y_i(\boldsymbol{x}_i^{\text{T}} \boldsymbol{\beta}) - \exp(\boldsymbol{x}_i^{\text{T}} \boldsymbol{\beta}) \Big] \\
&= \ell_c(\boldsymbol{\alpha}; \boldsymbol{r}) + \ell_c(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{r}).
\end{aligned} \tag{1.12}$$

Then, the E-Steps and M-Steps are:

1. *E-Step* - Given the current estimates of $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\alpha}^{(t)}$, compute $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$. Since $\ell_c(\cdot)$ is linear in the r_i's, this step can be simplified to updating the posterior memberships via Bayes Rule:

$$r_i^{(t+1)} = \mathbb{P}(R_i = 1 | Y_i = y_i; \boldsymbol{\theta}^{(t)}) \tag{1.13}$$

   Then, $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \ell_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{r}^{(t+1)})$.

2. *M-Step* - Maximize $\ell_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{r}^{(t+1)})$ by:

   - Maximize $\ell_c(\boldsymbol{\alpha}; \boldsymbol{r}^{(t+1)})$, which is equivalent to running a logistic regression of $\boldsymbol{r}^{(t+1)}$ on $\boldsymbol{W}$, where $\boldsymbol{W} = (\boldsymbol{w}_1^{\text{T}}, \ldots, \boldsymbol{w}_n^{\text{T}})^{\text{T}}$.
   - Maximize $\ell_c(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{r}^{(t+1)})$, which is equivalent to running a weighted Poisson regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ with weights $\boldsymbol{1} - \boldsymbol{r}^{(t+1)}$. Here, $\boldsymbol{X} = (\boldsymbol{x}_1^{\text{T}}, \ldots, \boldsymbol{x}_n^{\text{T}})^{\text{T}}$.

Since ZINB regression also requires estimating a heterogenity parameter, the optimization can be broken into two conditional maximization steps via an expectation-conditional-maximization (ECM) algorithm [28]. Thus, we perform iterative estimation of the heterogenity parameter and the regression parameters, such that the conditioning on the former allows the latter to be estimated via fitting a GLM. Details for an ECM algorithm in a ZINB regression model are given in the Appendix.

Inference

Testing of regression coefficients for predictors in a fully parameteric ZI count regression models is typically based on the asymptotic normality of the MLEs. Such testing applies to predictors in either the count regression component or the proportion of zero-inflation. Using the approximate standard errors, it is then straightforward to calculate Wald-based confidence intervals.

While tests of predictors are important, score tests on the zero-inflation structure in ZI count regression models have also been given considerable attention.
In particular,

$$\begin{aligned} H_0&: \pi = 0 \\ H_1&: 0 < \pi < 1, \end{aligned} \tag{1.14}$$

is employed to test the null hypothesis of a count regression model against the alternative hypothesis of a ZI count regression model. Many score tests pertaining to (1.14) have been developed in the literature. [29] developed a score test for zero-inflation in the Poisson setting, but where the zero-inflation probabilities are not a function of covariates. [30] extended this score test to the setting where the zero-inflation probabilities could depend on predictors. Similar score tests were developed in the negative binomial setting by [31].

Another test of interest is

$$\begin{aligned} H_0&: \theta = 0 \\ H_1&: \theta > 0, \end{aligned} \tag{1.15}$$

which is used to test for the presence of overdispersion in the ZI count regression models. Specifically, the null distribution is the ZIP regression model and the alternative distribution is the ZINB regression model. [32] developed a score test for testing the hypothesis (1.15). [33] provided a more comprehensive approach by developing score tests for each of the hypotheses in (1.14) and (1.15), as well as for testing both of them simultaneously.

[34] highlighted that since the null hypothesis for the tests in (1.14) and (1.15) are on the boundary of the parameter space, the standard asymptotic χ^2_1 distribution is conservative. An alternative is to employ a boundary likelihood ratio test using a modified χ^2 distribution [22]. The corresponding test statistic is characterized as having a limiting distribution that is a mixture of χ^2 distributions:

$$.5\chi^2_0 + .5\chi^2_1, \tag{1.16}$$

where χ^2_0 is a degenerate distribution with all its mass placed at 0. The distribution in (1.16) is also called a *chi-bar-squared distribution* when it pertains to the specific testing paradigm of uni-component versus two-component mixture models with known component densities [35].

The Vuong non-nested test [36] is also commonly used for testing the hypotheses in (1.14) and (1.15). However, [37] pointed out that by Vuong's definition, nesting occurs on the boundary, so a model is not strictly nested in its ZI counterpart. This does not imply that the models are non-nested and, hence, score tests or the boundary LRT should be used instead of the Vuong non-nested test. Finally, as noted in [22], model selction criteria, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), can be utilized to select between all the models discussed this far.

Various pseudo-R^2 measures for assessing goodness-of-fit of ZI count regression models were developed by [4]. One such measure the authors developed includes an adjustment to reward parsimony due to the chi-bar-squared limiting distribution of the boundary LRT. The adjusted-R^2 quantity for the ZIP regression model is given by

$$R^2_{\text{ZIP,adj}} = 1 - \frac{\ell_1(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{y}) - \ell_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}; \boldsymbol{y}) + p + q + 0.5}{\ell_1(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{y}) - \ell_1(\bar{y}\mathbf{1}_n, \mathbf{0}_n; \boldsymbol{y})}, \tag{1.17}$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ are the MLEs under the full ZIP regression model, and $\boldsymbol{z} = (z_1, \ldots, z_n)$ with $z_i = I\{y_i = 0\}$. Here, $\ell_1(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{y})$ is similar idea to the saturated likelihood in the GLM setting. Similarly, the adjusted-R^2 quantity for the ZINB regression model is given by

$$R^2_{\text{ZINB,adj}} = 1 - \frac{\ell_2(\boldsymbol{y}, \boldsymbol{z}, \widehat{\theta}; \boldsymbol{y}) - \ell_2(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\theta}; \boldsymbol{y}) + p + q + 1.5}{\ell_2(\boldsymbol{y}, \boldsymbol{z}, \widehat{\theta}; \boldsymbol{y}) - \ell_2(\bar{y}\mathbf{1}_n, \mathbf{0}_n, \widehat{\theta}; \boldsymbol{y})}, \tag{1.18}$$

where $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\theta}$ are the MLEs under the full ZINB regression model. Note that in both expressions, the log-likelihoods evaluated at $\boldsymbol{\pi} = 0$ (or $\boldsymbol{\alpha} = 0$) are simply the log-likelihoods for the corresponding non-ZI regression model. The above quantities are similar to the adjusted-R^2 formulas for non-ZI count regression models, as presented in [38].

Residual diagnostics for GLMs are typically employed when fitting ZI regression models. For example, one can assess Pearson, deviance, or Anscombe residuals for goodness of fit. However, goodness of fit is more difficult to define because we must examine a lack of fit in two processes, one of which is unobservable. Moreover, there

isn't a clear definition of fitted values in a zero-inflated regression model. It can be shown that the mean of a ZIP random variable is $(1-\pi_i)\mu_i$, where $\widehat{\pi}_i$ and $\widehat{\mu}_i$ are estimates of the mixing proportions and mean count for the, i$^{\text{th}}$ observation, respectively. Thus, the aforementioned definition of fitted value is the mean for the count distribution weighted by the probability of belonging in the count component. Another possible definition for fitted value is

$$\widehat{y}_i = \begin{cases} 0 & \text{if } R_i \geq .5 \\ \widehat{\mu}_i & \text{if } R_i < .5 \end{cases}, \tag{1.19}$$

where R_i is the posterior class probabilities and $\widehat{\mu}_i$ is again the mean of the count distribution for the i$^{\text{th}}$ subject. Thus, first perform a soft classification to the mixture componenets, and then take the mean of that component for prediction.

More recently, [39] and [34] have highlighted the utility of using randomized quantile residuals [40] for assessing the fit of ZI regression models. The definition from [40] is as follows: Let $F(y;\widehat{\mu}_i,\widehat{\phi})$ be a distribution function with estimated location $\widehat{\mu}_i$ and estimated nuissance parameters $\widehat{\phi}$. Let $a_i = \lim_{y\uparrow y_i} F(y;\widehat{\mu}_i,\widehat{\phi})$ and $b_i = F(y_i;\widehat{\mu}_i,\widehat{\phi})$. Then, the randomized quantile residual for y_i is

$$r_{q,i} = \Phi^{-1}(u_i) \tag{1.20}$$

where u_i is a uniform random variable on the interval $(a_i, b_i]$, and Φ^{-1} is the inverse cdf of a standard normal distributon. Then, the $r_{q,i}$ are exactly standard normal, apart from sampling variability in $\widehat{\mu}_i$ and $\widehat{\phi}$ [40]. The randomized quanitle residuals employ a similar idea to jittering, where the goal is improve data visualization by preventing an abundance of overlapping data points.

1.4 Bayesian ZI Models

Bayesian approaches for analyzing ZI count regression models have received increasing attention in the literature. [41] is one of the earliest papers where such a Bayesian analysis is performed. The paper presented a Bayesian hierarchical ZIP regression model that simultaneously models covariates and correlated count data. The approach was applied to count data on the efficacy of pesticides in controlling the reproduction of whiteflies. [42] presented the ZI power series (ZIPS) regression model, which provides a generalized setting for the ZIP and ZINB regression models. To define the power series distribution, let $b_0, b_1, b_2, \ldots$ be a sequence of nonnegative real

numbers. The partial sum of order $n \in \mathbb{N}$ is given by $g_n(\nu) = \sum_{k=1}^{n} b_k \nu^k$, $\nu \in \mathbb{R}$. The power series g is defined by $g(\nu) = \lim_{n\to\infty} g_n(\nu)$, and is denoted by $g(\nu) = \sum_{n=0}^{\infty} b_n \nu^n$. Letting $r \geq 0$ denote the radius of convergence of this series, the pmf of the power series distribution is given by

$$p(y;\nu) = \frac{b_y \nu^y}{g(\nu)}\ , \quad y \in \mathbb{N},\ 0 \leq \nu \leq r. \tag{1.21}$$

For Bayesian fitting of the ZIPS regression model, [42] assume that the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are *a priori* independent and specify multivariate normal priors with a scaled identity matrix as the variance-covariance matrix. The authors present their MCMC algorithm for generating samples from the respective full conditional distributions. They also provide their code, which is written in `WinBUGS` [43]. [44] had a similar setup as [42], but focused strictly on performing a Bayesian analysis of ZIP and ZINB regression models using a power prior as an informative prior. Their Bayesian approach was used to analyze data on road safety countermeasures.

The Bayesian ZIP regression model is typically formulated as $\text{logit}(\pi_i) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha}$ and $\log(\mu_i) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent. Moreover, the priors are usually "non-informative":

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \quad \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}), \tag{1.22}$$

where typically it is assumed that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \sigma_1 \boldsymbol{I}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \sigma_2 \boldsymbol{I}$. In the case of ZINB regression, common priors for the overdispersion parameter are uniform, gamma, and inverse-gamma.

Review of Bayesian Inference

Inference in the Bayesian setting is conducted through the posterior distribution of the parameters given the data. Let $\boldsymbol{X}$ be the data and $\boldsymbol{\theta}$ be the parameters. Let $\boldsymbol{X}|\boldsymbol{\theta} \sim f(\boldsymbol{X}|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$. Note, that $\pi(\boldsymbol{\theta})$ could depend on additional (random) parameters, i.e. *hyperparameters*, but we will assume that these are known.
Then, the posterior distribution of $\boldsymbol{\theta}|\boldsymbol{X}$ is given by

$$g(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{\pi(\boldsymbol{\theta}) f(\boldsymbol{X}|\boldsymbol{\theta})}{\int f(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{1.23}$$

Common estimates of $\boldsymbol{\theta}$ are the mean, median, and mode of the posterior distribution. Typically, outside of trivial cases, the posterior density $g(\cdot)$ cannot be calculated in

closed form. Therefore, inferences are based on a large chain of draws from the posterior distribution using MCMC algorithms. A frequently employed MCMC algorithm for sampling from the posterior distribution is *Gibbs Sampling*.

Gibbs Sampling

The key idea behind Gibbs sampling is that for a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$, it is easier to sample from several (univariate) conditional distributions, rather than trying to sample from the joint distribution by computing a normalization constant. Denote the k$^{\text{th}}$ sample by $\boldsymbol{X}^{(k)} = (X_1^{(k)}, \ldots, X_p^{(k)})^{\mathrm{T}}$. Start with an initial value $\boldsymbol{X}^{(0)}$, which is typically randomly chosen.
Then, for $k = 1, \ldots, N$, where N is sufficiently large :

- For $j = 1, \ldots, p$, update $x_j^{(k)}$ by sampling from the conditonal distribution

$$f(x_j^{(k)} | x_1^{(k)}, \ldots, x_{j-1}^{(k)}, x_{j+1}^{(k-1)}, \ldots, x_p^{(k-1)}). \tag{1.24}$$

  Note that the samples for $x_j^{(k)}$ are conditioned on the new samples for $x_1^{(k)}, \ldots, x_{j-1}^{(k)}$ (i.e. for $l = 1, \ldots j-1$), and the previous iteration samples for $x_{j+1}^{(k-1)}, \ldots, x_p^{(k-1)}$ (i.e. for $l = j+1, \ldots, p$).

- Repeat process until the desired number of samples are obtained.

It can be shown that these samples from the conditional distribution form an approximate sample from the joint distribution [45]. Moreover, it is common to discard the first 1000 samples in what is called *burn-in* period since the stationary distribution (i.e. the desired joint distribution) is not yet reached. Furthermore, it is common after the *burn-in* period to only keep every 20$^{\text{th}}$ or 100$^{\text{th}}$ sample since the sequential samples are correlated. Convergence of the MCMC algorithm is typically dictated by the traceplot of the samples for each $x_j^{(k)}$ (after the burn-in period), in addition to other numerical measures. Lastly, it is common to run multiple parallel chains with different initial values to ensure that all chains are converging to the same stationary distribution.

For the Bayesian ZIP Regression Model written in (1.21), the posterior complete-

data likelihood is

$$
\begin{aligned}
g(\boldsymbol{\beta}, \boldsymbol{\alpha} | \boldsymbol{y}, \boldsymbol{r}) &\propto \pi(\boldsymbol{\beta}, \boldsymbol{\alpha}) f(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{X}) f(\boldsymbol{r} | \boldsymbol{\alpha}) \\
&= \Big(\prod_{k=1}^{p} \phi(\beta_j | \sigma_\beta^2) \Big) \Big(\prod_{l=1}^{q} \phi(\alpha_l | \sigma_\alpha^2) \Big) \Big(\prod_{i=1}^{n} \pi_i(\boldsymbol{\alpha})^{r_i} (1 - \pi_i(\boldsymbol{\alpha}))^{1 - r_i} \Big) \\
&\times \Big(\prod_{j=1}^{n} (1 - r_i) \exp(-\mu_i(\boldsymbol{\beta})) \mu_i(\boldsymbol{\beta})^{y_i} \Big) \\
&= \Big(\prod_{l=1}^{q} \phi(\alpha_l | \sigma_\alpha^2) \prod_{i=1}^{n} \pi_i(\boldsymbol{\alpha})^{r_i} (1 - \pi_i(\boldsymbol{\alpha}))^{1 - r_i} \Big) \\
&\times \Big(\prod_{j=1}^{p} \phi(\beta_j | \sigma_\beta^2) \prod_{j=1}^{n} (1 - r_i) \exp(-\mu_i(\boldsymbol{\beta})) \mu_i(\boldsymbol{\beta})^{y_i} \Big) \\
&= g(\boldsymbol{\alpha} | \sigma_\alpha^2, \boldsymbol{r}) g(\boldsymbol{\beta} | \sigma_\beta^2, \boldsymbol{y}, \boldsymbol{r}),
\end{aligned}
\tag{1.25}
$$

where $\phi(\cdot)$ the Gaussian density with variance σ^2. Note that the posterior factors into two condtionals for the mixing proportions and the mean of the Poisson state. This can simplify the Gibbs sampler significantly. Note that the likelihood depends on the latent variable $\boldsymbol{r}$. Both [42] and [46] incorperate data augmentation into their MCMC samplers. Here, we assume that σ_α^2 and σ_β^2 are known.

A summary of the process is as follows :

1. Begin with intial values of $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\alpha}^{(0)}$.

2. For $t = 1, 2, \ldots, N$, do:

   a) For $i = 1, \ldots, n$, generate $R_i^{(t)}$ with

$$
\mathbb{P}(R_i^{(t)} = 1) = \begin{cases} \frac{\pi_i(\boldsymbol{\alpha}^{(t-1)})}{\pi_i(\boldsymbol{\alpha}^{(t-1)}) + (1 - \pi_i(\boldsymbol{\alpha}^{(t-1)})) \exp(-\mu_i(\boldsymbol{\beta}^{(t-1)}))} & y_i = 0 \\ 0 & y_i > 0. \end{cases}
\tag{1.26}
$$

   b) Generate $\boldsymbol{\alpha}^{(t)}$ from $g(\boldsymbol{\alpha} | \sigma_\alpha^2, \boldsymbol{r}^{(t)})$.

   c) Generate $\boldsymbol{\beta}^{(t)}$ from $g(\boldsymbol{\beta} | \sigma_\beta^2, \boldsymbol{y}, \boldsymbol{r}^{(t)})$.

An example of a fitted Bayesian ZIP regression model can be seen in Section 1.6.

Bayesian Testing and Diagnostics

Bayesian approaches to test and construct influence diagnostics have also been proposed in the literature. [47] proposed a Bayes factor based on a suitable objective

prior for testing a Poisson regression model versus a ZIP regression model. Bayesian inference involving specific ZI count regression models has also been treated in the literature, including approaches for ZINB regression [25], ZIGP regression [48], and ZICMP regression [49]. In each paper, the authors present an MCMC sampler for the particular ZI model under consideration, followed by a discussion of Bayesian case influence diagnostics and relevant model selection criteria. For Bayesian influence diagnostics, the primary approach is based on *case-deletion*, where the impact of deleting an observation on the estimates is directly assessed by measures such as likelihood distance and Cook's distance. Some of the model criteria discussed include deviance information criterion (DIC) [50], the expected BIC (EPIC) [51], and the log-pseudo-marginal likelihood (LMPL) statistic. The LMPL statistic requires calculation of the conditional prediction ordinate (CPO) statistic of [52]. Let $\boldsymbol{Y}|\boldsymbol{\theta} \sim f(\boldsymbol{y}|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$. Let $\boldsymbol{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, y_n)^{\mathrm{T}}$ be the vector of responses with the i^{th} case deleted. The CPO for the i^{th} observation is defined as

$$\mathrm{CPO}_i := f(y_i|\boldsymbol{y}_{-i}).$$

In other words, the CPO estimates the likelihood of observing y_i conditional on observing $\boldsymbol{y}_{-i}$. Furthermore, note that

$$\begin{aligned}
\mathrm{CPO}_i &= f(y_i|\boldsymbol{y}_{-i}) \\
&= \frac{f(\boldsymbol{y}_{-i}|y_i)f(y_i)}{f(\boldsymbol{y}_{-i})} \\
&= \Big(\frac{f(\boldsymbol{y}_{-i})}{f(\boldsymbol{y})}\Big)^{-1} \\
&= \left(\frac{1}{f(\boldsymbol{y})}\int f(\boldsymbol{y}_{-i}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right)^{-1} \\
&= \left(\int (f(y_i|\boldsymbol{\theta}))^{-1}\frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{y})}d\boldsymbol{\theta}\right)^{-1} \\
&= \left[\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\big\{(f(y_i|\boldsymbol{\theta}))^{-1}\big\}\right]^{-1}.
\end{aligned}$$

The estimate of CPO_i is then defined as the inverse harmonic mean

$$\widehat{\mathrm{CPO}_i} = \left[N^{-1}\sum_{t=1}^{N}\big(f(y_i|\boldsymbol{\theta}^{(t)})\big)^{-1}\right],$$

where $\boldsymbol{\theta}^{(t)}$ denote Monte Carlo samples from the posterior distribution. Hence, larger values of $\widehat{\text{CPO}}_i$ indicate a better fit. Then, $LPML = \sum_{i=1}^{n} \log(\widehat{CPO}_i)$, such that larger values of LPML indicate a better fit.

Recent Bayesian Advances

[53] proposed a class of Bayesian Generalized Additive Models for ZI count responses in the Generalized Additive Models for Location, Shape, and Scale (GAMLSS) framework. Attractive features of their model include flexible modeling for all the parameters in a ZI regression model (including the scale or heterogenity parameter), as well as providing an efficient framework for extending to multilevel models such as spatial regression. For illustration, let $\{Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{V}_i\}$ be a random sample such that $Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i = \boldsymbol{z}_i, \boldsymbol{V}_i = \boldsymbol{v}_i \sim \text{ZINB}(\pi_i(\boldsymbol{z}_i), \mu_i(\boldsymbol{x}_i), \theta_i(\boldsymbol{v}_i))$.
Furthermore, suppose the following :

$$\begin{aligned} \text{logit}(\pi_i) &= \alpha_0 + f_1(\boldsymbol{z}_i) + \cdots + f_{J_\pi}(\boldsymbol{z}_i), \\ \log(\mu_i) &= \beta_0 + h_1(\boldsymbol{x}_i) + \cdots + h_{J_\mu}(\boldsymbol{x}_i), \\ \log(\theta_i) &= \gamma_0 + g_1(\boldsymbol{v}_i) + \cdots + g_{J_\theta}(\boldsymbol{v}_i), \end{aligned} \tag{1.27}$$

where we assume each f_j, h_k, and g_l can be approximated by linear combinations of basis functions. So, for example, each f_j can be expressed such that $f_j(\boldsymbol{z}_i) \approx \sum_{d_j=1}^{D_j} \alpha_{j,d_j} B_{j,d_j}(\boldsymbol{z}_i)$, where $B_{j,d_j}(\cdot)$ are basis functions (ex: B-spline basis). Then, letting $\boldsymbol{\alpha}_j = \left(\alpha_{j1}, \ldots, \alpha_{jD_j}\right)^{\mathrm{T}}$, which is the vector of regression coefficients for $f_j(\cdot)$, we assign a multivariate Gaussian prior $\boldsymbol{\alpha}_j|\tau_j^2 \sim \mathcal{N}(\mathbf{0}_{D_j}, \tau_j^2 \boldsymbol{K}_j)$ to enforce smoothness conditions. Here, $\boldsymbol{K}_j$ is a prior (diagonal) precision matrix, and τ_j^2 is the hyperparameter for smoothing variance. We assume all the functions $h_l(\cdot)$ and $g_l(\cdot)$ can be approximated analogously. For details on IWLS implementation in a Metropolis-Hastings sampler, see [53].

[54] developed a Bayesian latent factor zero-inflated (LZIP) model for analyzing correlated zero-inflated counts, which was used to study moleclar differences among breast cancer patients. In their model, the formulation of a random variable is ZIP distributed when $Y \sim (1-\pi)\text{I}\{W = 0\} + \pi p(y|\mu)\text{I}\{W = 1\}$, where $p(y|\mu)$ is the Poisson mass function with mean μ, and W is a latent "at-risk" indicator such that $Y \sim 0$ with probability $1-\pi$ when $W = 0$, and Y is drawn from a Poisson distribution having mean μ with probability π when $W = 1$.

Then, using the complementary log-log link and log link,

$$\begin{aligned}\operatorname{cloglog}(\pi) &= \operatorname{cloglog}(\mathbb{P}(W=0)) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha},\\ \log(\mu) &= \log(E(Y|W=1)) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta},\end{aligned}$$

which then implies $\pi = 1 - \exp(-\exp(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha}))$. Therefore, π is equivalent to the probability that a Poisson random variable, call it Z_1, with mean $\mu_1 = \exp(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha})$, is bigger than zero. Hence, $W = 1$ if and only if $Z_1 > 0$. In other words, Z_1 is a latent variable indicating the capability of being at risk. Similarly, define another Poisson latent variable $Z_2 := (Y|Z_1 > 0)$ with mean $\mu_2 = \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$, which is the count conditional on the subject being at risk for the given outcome. Then, rewriting the ZIP model in terms of Z_1 and Z_2, the model is

$$\begin{aligned}&Y \sim (1-\pi)\mathrm{I}\{Z_1 = 0\} + \pi p(z_2|\mu_2)\mathrm{I}\{Z_1 > 0\},\\ &\pi = \mathbb{P}(W=1) = \mathbb{P}(Z_1 > 0) = 1 - \exp(-\exp(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha})),\\ &\mu_1 = \mathbb{E}(Z_1) = \exp(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha}),\\ &\mu_2 = \mathbb{E}(Z_2) = \mathbb{E}(Y|Z_1 > 0) = \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}).\end{aligned} \tag{1.28}$$

The variable Z_2 can be viewed as the "potential" count that would have been observed if the subject had been at risk.

It is reasonable to assume that Z_1 and Z_2 are positively correlated for most instances. For example, as [54] wrote, "in cancer genomics, we might expect patients with increased risk of pathway activation to also have more genes with CNVs given activation." This is similar to the idea behind the ZIP(τ) model in [3], where the zero-inflation probabilites are functionally related to the mean of the Poisson state. Then, to accommodate association between μ_1 and μ_2, we assume that μ_1 and μ_2 can be written as a multiplicative function of subject-specific latent factors, $\boldsymbol{\xi}$. Then, the resulting latent factor ZIP (LZIP) model is

$$Y|\boldsymbol{\xi} \sim (1-\pi)\mathrm{I}\{Z_1 = 0\} + \pi p(z_2|\mu_2)\mathrm{I}\{Z_1 > 0\}, \tag{1.29}$$

where

$$\begin{aligned}&\pi = \mathbb{P}(W=1|\boldsymbol{\xi}) = \mathbb{P}(Z_1 > 0|\boldsymbol{\xi}) = 1 - \exp(-\mu_1),\\ &\mu_1 = \mathbb{E}(Z_1|\boldsymbol{\xi}) = (\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\xi})\exp(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha}),\\ &\mu_2 = \mathbb{E}(Z_2|\boldsymbol{\xi}) = \mathbb{E}(Y|Z_1 > 0, \boldsymbol{\xi}) = (\boldsymbol{\lambda}_2^{\mathrm{T}}\boldsymbol{\xi})^{\mathrm{T}}\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}),\end{aligned} \tag{1.30}$$

and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)^{\mathrm{T}}$ is an $L \times 1$ vector of subject-specific latent factors with $\xi_l > 0$

for all $l = 1, \ldots, L$ to ensure $\mu_k > 0$ $(k = 1, 2)$, and $\lambda_k = (\lambda_{k1}, \ldots, \lambda_{kL})^{\mathrm{T}}$ is an $L \times 1$ vector of loadings for the k$^{\text{th}}$ component. Again, we assume $\lambda_{kl} > 0$ for all k, l. [54] discusses some strategies for choosing L, such as the *widely applicable information criterion* (WAIC) [55]. Here, the latent factor, $\boldsymbol{\xi}$, accounts for "between-subject heterogeneity potentially due to unmeasured subject-level confounding". [54] assumes that $\xi_l \overset{\text{iid}}{\sim} \text{Gamma}(\gamma, \gamma)$, where the authors recommend setting $\gamma = 1$.

Since $\mathbb{E}(\xi_l) = 1$, it follows the population-averaged mean of Z_k is

$$\mathbb{E}(Z_k) = \begin{cases} \left[\sum_{l=1}^{L} \lambda_{1l}\right] \exp(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha}) = \exp(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha} + v_1) & \text{if } k = 1 \\ \left[\sum_{l=1}^{L} \lambda_{2l}\right] \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) = \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + v_2) & \text{if } k = 2, \end{cases} \tag{1.31}$$

where $v_k = \log(\sum_{l=1}^{L} \lambda_{kl})$. Thus, v_k is similar to a population-averaged intecept for the k$^{\text{th}}$ component. [54] also extends the LZIP model to multiple zero-inflated outcomes, such as in longitudinal studies.

In [54], $\lambda_{kl} \overset{\text{iid}}{\sim} \text{Gamma}(a, b)$ with $a = b = .001$. For categorical predictors, w_h or x_h, $\exp(\alpha_h)$ or $\exp(\beta_h)$ are assigned $\text{Gamma}(c, d)$ priors, with $c = d = .001$, since these priors are conjugate for the model. Then, for continuous predictors, α_h and β_h are assigned uninformative normal priors. For details on the Gibbs sampler with data augmentation, see [54].

[56] also discusses Bayesian longitudinal modeling via random effects. In the article, the authors discuss repeated measures modeling for the Poisson hurdle regression model, ZIP regression model, and the *zero-altered* model. See Section 1.10 for definitions of hurdle models and "zero-altered" models. For the random effect on the i$^{\text{th}}$ subject, $\boldsymbol{b}_i = (b_{i1}, b_{i2})^{\mathrm{T}}$, it is assumed that $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where b_{i1} and b_{i2} represent random intercepts for the zero-state and the Poisson mean, respectively. Then, it is assumed that $\boldsymbol{\Sigma} \sim \text{I-W}(2, \boldsymbol{I}_2)$, where $\text{I-W}(2, \boldsymbol{I}_2)$ denotes a Inverse-Wishart distribution with 2 degrees of freedom and scale matrix $\boldsymbol{I}_2$. The authors also consider $\boldsymbol{\Sigma}$ being distributed with a product normal distribution [57]. [56] then discusses a Metropolis-Hastings algorithm with iteratively weighted least squares proposal densities for sampling from the posterior distribution, as well as model selection criteria such as DIC and the CPO statistic.

It was noted by [58] that the ZIP and ZINB regression models can provide unsatisfactory fits when there is extreme incidence of zeros (above 80%); in particular, those models can be unable to find important covariates. To mitigate this issue, the authors propose a k-ZIG regression model, which allows for more flexible modeling between the zero and count components. In particular, suppose $G(y|\Theta)$ is a zero-inflated dis-

tribution with a point mass at zero and $G_0(y|\Theta_0)$ is a pmf that is non-degenerate. In other words, $G(y|\Theta) = q\mathrm{I}\{y = 0\} + (1 - q)G_0(y|\Theta_0)$. Then, inputting $G(\cdot)$ as the "non-degenerate" mass function into a ZI mass function, we obtain

$$\begin{aligned}\pi(y|p, q, \Theta_0) &= p\mathrm{I}\{y = 0\} + (1 - p)G(y|\Theta) \\ &= p\mathrm{I}\{y = 0\} + (1 - p)\big[q\mathrm{I}\{y = 0\} + (1 - q)G_0(y|\Theta_0)\big] \\ &= \big(p + (1 - p)q\big)\mathrm{I}\{y = 0\} + (1 - p)(1 - q)G_0(y|\Theta_0).\end{aligned} \tag{1.32}$$

In other words, to account for highly excessive zeros, we assume the density of the data to be a mixture of a point mass at zero and a ZIP density. The allowance of three sources of zeroes can make this model more desirable at explaining heavily zero-inflated data. It can be shown that p and q are not identifiable in (1.30), and so we reparameterize by assuming $(1 - p) = (1 - \theta)^{k-1}$ and $(1 - q) = (1 - \theta)$, which yields

$$\pi(y|\theta, \Theta_0, k) = (1 - (1 - \theta)^k)\mathrm{I}\{y = 0\} + (1 - \theta)^k G_0(y|\Theta_0). \tag{1.33}$$

The above density in (1.31) is, thus, called the k-ZIG(θ, Θ_0) distribution. The parameter k increases (decreases) the amount of zero-inflation. See [58] for further details on regression for the k-ZIG(θ, Θ_0) distribution, prior selection, and MCMC sampling.

We end this section by highlighting a few more diverse problems with zero-inflated data that were addressed from using Bayesian methods. [59] modeled the ordinal outcomes of smoking and chewing tobacco jointly by developing a bivariate zero-inflated probit regression model (ZIBOP). This is necessary since it is common for non-smokers to also not chew tobacco (i.e. joint zero-inflation of the bivariate response). [60] developed a Bayesian multivariate measurement error model with zero-inflation. The authors applied their measurment error model to a National Health and Nutrition Examination Survey (NHANES) data set to model the the amount of 12 food groups (vegtables,fruit,oil,etc.) consumed in a 24 hour period regressed on age, gender, and race. The authors noted that it is common for some dietary groups to be consumed daily by almost everyone, while other food groups are episodically consumed (i.e. zero-inflation). Furthermore, the dietary intake can exhibit sizeable measurement error and consumption of certain food groups are correlated. [61] also developed a Bayesian trivariate ZIP regression model to predict the number of third-party liability automobile claims , motor collison claims, and other claims for automobile insurance. The authors developed an MCMC algorithm for analyzing this model.

Table 1.1: Results for estimating simulated data from a ZIP regression model using the two R functions in the pscl, glmmTMB, and VGAM packages, and the three SAS procedures, GENMOD, NLMIXED, and COUNTREG. Notice the highly variable results across the five methods for $\boldsymbol{\alpha}$ when $n = 50$.

n	Procedure	β_0 (= 3.000)	β_1 (= −1.500)	α_0 (= 0.500)	α_1 (= −0.500)
50	zeroinfl	2.998 (0.230)	-1.511 (0.240)	1.108 (10.197)	-1.387 (15.001)
	vglm	2.997 (0.229)	-1.506 (0.239)	0.844 (2.228)	-1.030 (3.874)
	glmmTMB	2.992 (0.213)	-1.511 (0.229)	3.879 (62.870)	-4.295 (70.208)
	GENMOD	2.998 (0.230)	-1.512 (0.240)	0.758 (1.203)	-0.852 (1.333)
	NLMIXED	2.998 (0.230)	-1.511 (0.240)	1.038 (5.681)	-1.307 (8.667)
	COUNTREG	2.998 (0.230)	-1.511 (0.240)	6.453 (262.742)	-11.971 (371.527)
250	zeroinfl	2.996 (0.084)	-1.500 (0.098)	0.523 (0.375)	-0.532 (0.356)
	vglm	2.996 (0.084)	-1.500 (0.098)	0.522 (0.375)	-0.532 (0.355)
	glmmTMB	3.002 (0.081)	-1.506 (0.092)	0.536 (0.368)	-0.550 (0.351)
	GENMOD	2.996 (0.084)	-1.500 (0.098)	0.523 (0.375)	-0.532 (0.355)
	NLMIXED	2.996 (0.084)	-1.500 (0.098)	0.523 (0.375)	-0.532 (0.356)
	COUNTREG	2.996 (0.084)	-1.500 (0.098)	0.523 (0.375)	-0.532 (0.356)

1.5 Software and Numerical Demonstrations

Many statistical software programs have routines for estimating ZI count regression models, but the scope of such functions is usually limited to estimating ZIP and ZINB regression models. In SAS [62], three available procedures are PROC GENMOD, PROC NLMIXED, and PROC COUNTREG. In R [63], two of the major functions available are zeroinfl and vglm, which are within the pscl [64] and VGAM [65] packages, respectively. Recently, another R package, glmmTMB [66], was developed for count data regression, including zero-inflated regression models. glmmTMB employs similar syntax and interface to lme4 [67], where fixed and random effects can both be specified in the zero-inflation state and count component. For estimation, all of the aforementioned functions employ gradient-based methods, such as Newton-Raphson or iteratively reweighted least squares (IRLS), by default. Mixed models in glmmTMB are estimated via optimization of the marginal likelihood function through Gauss-Hermite quadrature.

We demonstrate the accuracy of the estimates obtained using the aforementioned six procedures through a brief simulation study. We generated $B = 1000$ datasets of sizes $n \in \{50, 250\}$ from a ZIP regression model and ZINB regression model. The parameters for these models are given in the headers of Tables 1.1 and 1.2, respectively. We report the mean and standard deviation of the ZI count regression estimates using the different procedures such that all arguments are set to their respective defaults. The ZIP regression estimates in Table 1.1 are nearly identical for $\boldsymbol{\beta}$ in the different

Table 1.2: Results for estimating simulated data from a ZINB regression model using the two R functions in the pscl and VGAM packages, and the three SAS procedures, GENMOD, NLMIXED, and COUNTREG. Notice the considerably different results for COUNTREG.

n	Method	β_0 (= 3.000)	β_1 (= 1.200)	α_0 (= −0.500)	α_1 (= 0.500)	θ (= 4.482)
50	zeroinfl	2.986 (0.160)	1.168 (0.160)	-1.023 (0.563)	0.605 (0.536)	5.576 (1.221)
	vglm	2.986 (0.160)	1.168 (0.160)	-1.023 (0.563)	0.605 (0.536)	5.576 (1.221)
	glmmTMB	3.003 (0.209)	1.196 (0.084)	0.549 (0.648)	-0.536 (0.308)	5.187 (1.460)
	GENMOD	2.986 (0.160)	1.168 (0.160)	-1.023 (0.563)	0.605 (0.536)	5.576 (1.221)
	NLMIXED	2.986 (0.160)	1.168 (0.160)	-1.023 (0.563)	0.605 (0.536)	5.576 (1.221)
	COUNTREG	2.871 (0.304)	1.011 (0.305)	-0.360 (63.363)	-106.078 (473.480)	2.208 (2.121)
250	zeroinfl	3.003 (0.076)	1.189 (0.077)	-0.573 (0.217)	0.486 (0.217)	4.617 (0.412)
	vglm	3.004 (0.076)	1.189 (0.077)	-0.573 (0.217)	0.486 (0.217)	4.617 (0.412)
	glmmTMB	2.996 (0.093)	1.201 (0.036)	-0.524 (0.268)	0.514 (0.124)	4.603 (0.547)
	GENMOD	3.004 (0.076)	1.189 (0.077)	-0.573 (0.217)	0.486 (0.217)	4.617 (0.412)
	NLMIXED	3.004 (0.076)	1.189 (0.077)	-0.573 (0.217)	0.486 (0.217)	4.617 (0.412)
	COUNTREG	2.901 (0.216)	1.137 (0.134)	-1.844 (11.082)	-56.140 (533.695)	1.540 (2.398)

procedures, but with fairly noticeable differences in the estimates of $\boldsymbol{\alpha}$ when $n = 50$, especially for PROC COUNTREG. The ZINB regression estimates in Table 1.2 are nearly identical for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ with the different procedures, but there are sizeable differences in the estimates of $\boldsymbol{\alpha}$ when $n = 50$. However, this time PROC COUNTREG demonstrates quite different results for both sample sizes. These numerical results are consistent with those obtained in Liu et al. [68], who performed an extensive simulation study that addresses the performance of ZI estimation procedures in SAS and R.

We also performed a brief timing study for comparing the six procedures discussed above. We generated datasets of different sample sizes from a ZIP and ZINB regression model, and timed each of the five procedures using their default settings. PROC GENMOD was found to perform the quickest for nearly all of the settings considered. The zeroinfl function typically took the longest when estimating the ZIP regression model, while the vglm function typically took the longest when estimating the ZINB regression model. More details, including the actual timing results, are given in the Appendix.

We note that the gamlss package [69] can also estimate ZIP and ZINB regression models, as well as some related models discussed in Section 1.10, using the GAMLSS framework. Estimation is performed using a maximum penalized likelihood approach, which differs from the pscl and VGAM packages. Thus, we did not include a comparison with estimates obtained using the gamlss package.

Other statistical software have built-in routines to estimate ZIP and ZINB regression models. To estimate these models in Mplus [70], place the (i) option after the count response variable in the count statement. In the Stata software [71], the

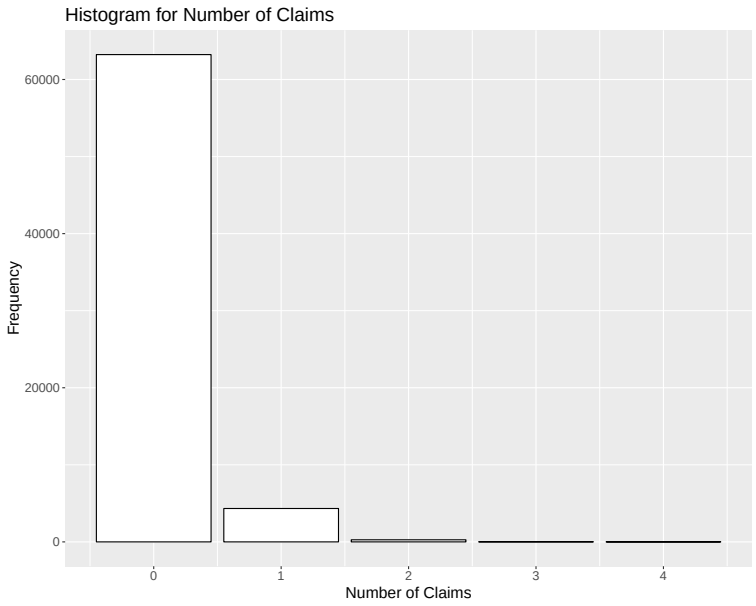


Figure 1.2: Histogram of the Number of Claims

`zip` and `zinb` functions can be used, respectively, to estimate ZIP and ZINB regression models. Also, the NCSS software [72] has routines for estimating both of these models, which is found under the *Regression with Count Data* menu.

1.6 Example: Insurance Data

The data set consists of 67856 one-year automobile insurance policies in Australia for the years 2004 and 2005. The response is the number of claims on a policy with predictors associated with the policyholder and the vehicular characteristics. The policyholder variables are gender, area of residence (A,B,C,D,E,F), and the age of the driver (1,2,3,4,5,6), where 1 denotes the youngest group and 6 denotes the oldest group. The vehicular variables are vehicle value (in $10,000s), vehicle body (see Table 1.3), and the age of the vehicle (1,2,3,4). Here, higher values of age represent older vehicles. All predictors are categorical except for vehicle value. Moreover, the natural log of the length of exposure, which is the length of time of the policy in years, is utililized as an offset in the count regression component. The data set is from the textbook *Generalized Linear Models for Insurance Data* [73], and can be accessed in the *dataCar* data set in the `insuranceData` `R` package [74].

EDA and Initial Variable Selection

A histogram of the number of claims can be seen in Figure 1.2. Overall, 93.19% of the policies have no claims, and an additional 6.9% had one claim. Vehicle body was binned with the goal of grouping similar vehicles (see Table 1.3). The vehicle group of *general* was taken to be the reference category. Initial variable selection for the count component was based on backwards elimination in the Poisson GLM of the aforementioned variables. A list of variables selected for the count component can be seen in Table 1.4. For the zero-inlation component, variables were chosen in a forward stepwise manner. After vehicle value was selected, no other significant predictors were found for the zero-inflation component.

The Poisson, negative binomial, generalized Poisson (GP), and Conway-Maxwell-Poisson (CMP), along with their zero-inflated counterparts were fit to the data. First, we performed the boundary LRTs discussed in Section 1.3. The test of zero-inflation for the Poisson regression setting and negative binomial regression settings have test statistics of 69.8893 (p-value = 3.13×10^{-17}) and 30.5087 (p-value = 1.66×10^{-8}), respectively. Thus, with respect to the Poisson and negative binomial model, we can conclude that the zero-inflated counterparts provide a better fit. Moreover, using the boundary LRT to compare ZIP versus ZINB, the test statistic is 0.9746 (p-value = 0.08), and therefore, we cannot conclude that there is evidence of overdispersion relative to the ZIP model.

The AIC and BIC statisitcs can be seen in Table 1.5. In general, we can see that the zero-inflated counterparts are improvements over the single component of a count regression. Moreover, we see that by both criteria, the ZIP model provides the best fit to the data.

[75] provides rules of thumbs for interpreting model differences in BIC, which is based on a Bayes factor. According to [75], a difference in BIC between models M_1 and M_2 that is between 0-2 is weak evidence of superiority, 2-6 is positive evidence

Table 1.3: Bins of Vehicle Body Types

Bins	**Original Body Types**
Convertibles	Hardtop,Convertibles
Vans	Caravan,Panelvan
Two-Seaters	Roadster,Coupe
Bus	Bus
Utility	Utility
General	Hatchback,Station Wagon,Sedan,Truck

of superiority, 6-10 is strong evidence of supremacy, and finally > 10 is very strong evidence of supremacy. From these rules of thumb, we conclude that ZIP is at least "strongly" superior to every other considered model, the other zero-inflated models (ZINB, ZIGP, ZICMP) are similar, and that all non-zero-inflated models are considerably weaker than the zero-inflated counterparts. [76] offer similar criteria for AIC differences: a difference in AIC of less than 2 suggests a non-substantial difference, between 3 and 7 is a substantial difference, and > 10 indicates a large difference. So, according to [76], all the zero-inflated models are roughly equivalent, whereas the non-zero-inflated counterparts are substantially less probable models. [76] also discusses Akaike weights and evidence ratios as a means to compare competing models.

The randomized quantile residuals (see Figure 1.3) were then examined for each model. Note that the zero-inflated counterparts typically produce better fits with respect to the quantiles of the $\mathcal{N}(0, 1)$ line, although all models deviate from the line for the extremely rare larger counts; i.e., when the number of counts is 3 or 4.

Finally, we fit a Bayesian ZIP model discussed in Section 1.4. The priors were taken to be fairly non-informative with $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 10\boldsymbol{I})$ and $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, 10\boldsymbol{I})$. The MCMC sampler was constructed in `JAGS` [77] using the zeros trick, and the code is

Table 1.4: Count Component Variables Selected

Variables Selected
Vehicle Value
Convertible Indicator
Bus Indicator
Van Indicator
Two-Seater Indicator
Area D Indicator
Age of Driver

Table 1.5: BIC and AIC Values for Fitted Models

Model	**BIC**	**AIC**
Poisson	34934.4249	34824.9231
NB	34905.1512	34786.5243
GP	34916.9900	34798.3600
CMP	34905.5000	34786.8700
ZIP	34890.9244	34763.1724
ZINB	34900.7921	34763.9149
ZIGP	34901.9626	34765.0855
ZICMP	34901.2733	34764.3961

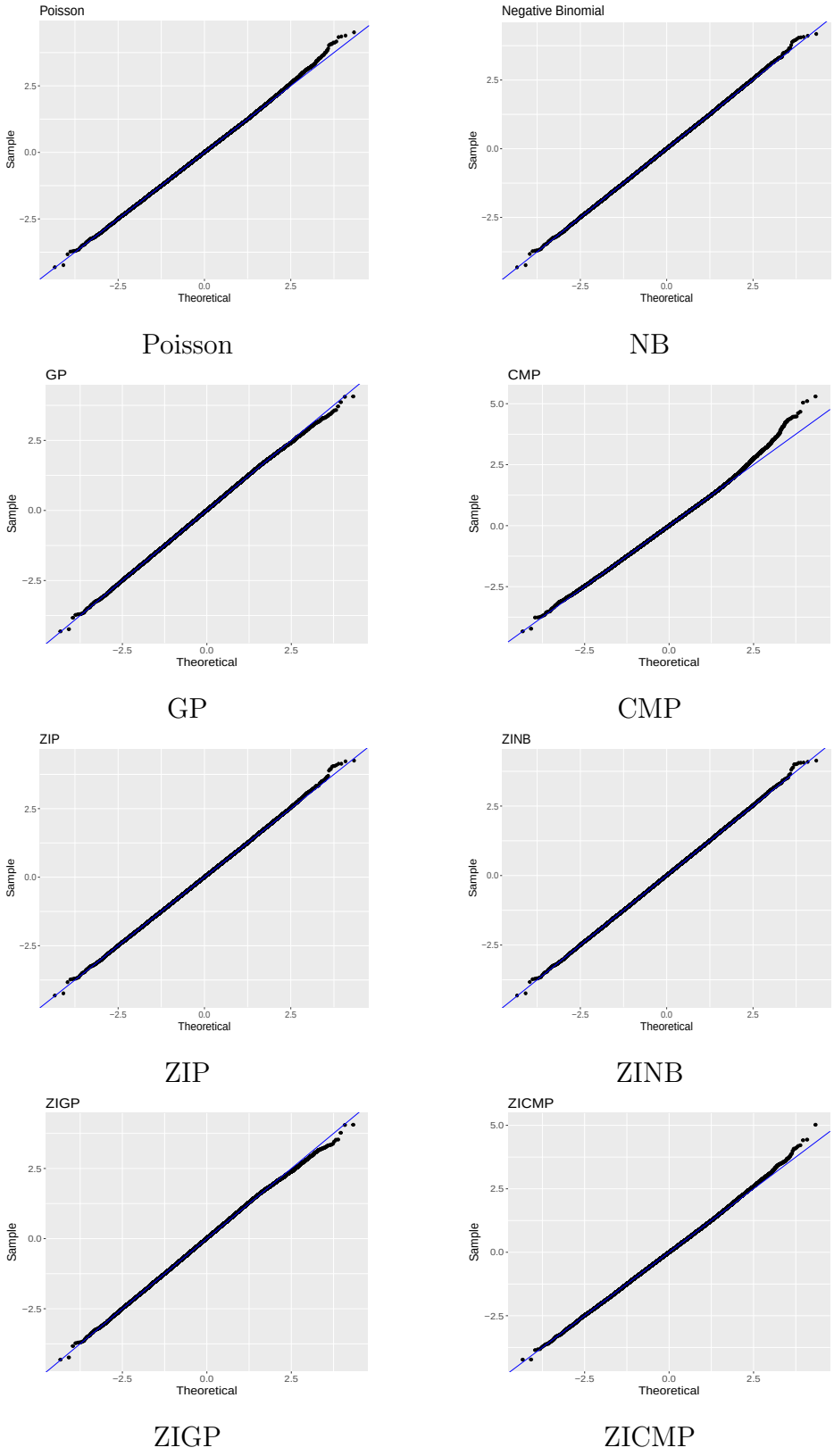


Figure 1.3: Randomized Quantile Residual Plots

included in the Appendix. Table 1.6 shows the coefficients from the frequentist and Bayesian ZIP regressions. The MCMC sampler was ran for 6000 iterations, and did not include a burn-in.

The covariate effects indicate that higher age groups typically file less claims than younger age groups. Moreover, two seaters, vans, convertibles, and buses generally result in higher claim counts than the general cars or utility vehicles. The effect due to vehicle value on average counts is a little harder to interpret directly since it is a covariate in both the Poisson and zero-inflation state. But, we can see the sign for the log-odds coefficient in the zero-inflation state for vehicle value is negative, which means that lower costs cars are more likely to experience zero-inflation compared to higher cost cars. In other words, it is more probable for lower cost vehicles to exhibit zero claims in contrast to higher cost vehicles. This could be because lower cost automobiles are more likely to have smaller claim amounts that are below the deductible, and therefore, will not be filed.

1.7 ZI Count Regression Models for Handling Data Dispersion

Relative to the Poisson distribution, many count data sets are heavily right-skewed and exhibit excess zero observations. As noted in [78] and [39], overdispersion has

Table 1.6: ZIP Model Coefficients

Variable	**Frequentist**	**Bayes**
Count Component		
Intercept	-1.108	-1.109
Vehicle Value	-0.066	-0.067
Bus Ind	.861	.809
Convertible Ind	0.027	0.024
Van Ind	0.133	0.127
Two-Seater Ind	0.419	0.412
Utility Ind	-0.219	-0.221
Area D Ind	-0.127	-0.128
Age Group 2	-0.178	-0.177
Age Group 3	-0.235	-0.234
Age Group 4	-0.259	-0.258
Age Group 5	-0.473	-0.472
Age Group 6	-0.446	-0.472
Zero Component		
Intercept	0.183	0.189
Vehicle Value	-0.664	-0.683

the tendacy to increase the proportion of zeros such that other distributions, like the negative binomial, can improve the fit. However, even better fits can be obtained through overdispersed models that simultaneously characterize excess zeros.

When the negative binomial still fails to provide a good fit to the data, the generalized Poisson distribution [79, 80] can often provide an improved fit. For two given parameters, $\mu > 0$ and $\max\{-1, -\mu/4\} \leq \alpha < 1$, one parameterization of the generalized Poisson pmf [81] is

$$p(y; \mu, \alpha) = \begin{cases} \mu(\mu + \alpha y)^{y-1} \exp\{-(\mu + \alpha y)\}/y! & y \in \mathbb{N} \\ 0 & \text{if } y > m \text{ when } \alpha < 0, \end{cases} \tag{1.34}$$

where m is the largest positive integer such that $\alpha + m\mu > 0$, when the dispersion parameter α is negative. When $\alpha = 0$, the above reduces to the Poisson pmf (equadispersion), and $\alpha > 0$ and $\alpha < 0$ represent count data with overdispersion and underdispersion, respectively.

[82] and [78] were the earliest works to study the ZI generalized Poisson (ZIGP), where the mean μ is related to the covariates $\boldsymbol{x}$ through the log link function. The former also established the consistency and asymptotic normality of the MLEs for the parameters in the ZIGP regression model. [83] developed a score test to determine whether the ZIGP regression model is necessary over the ZIP or ZINB regression models. [84] provdied an extension of the ZIGP regression model that allows the dispersion parameter to be related to a vector of covariates. Computational routines for this model were made available in the `R` package `ZIGP`, which is archived as of July 2017. However, the ZIGP regression model can be fit in the `glmmTMB` package [66]. Applications where the ZIGP regression model has been demonstrated to provide a better fit compared to the ZIP and ZINB regression models are data on domestic violence occurances [78], outsourcing of patent filing process [84], and mapping quantitative trait loci [85].

The Conway-Maxwell-Poisson (CMP) distribution of [86] is another flexible distribution for count data expressing overdispersion or underdispersion [39].
This two-parameter distribution has pmf

$$p(y; \mu, \nu) = \frac{\mu^y}{(y!)^\nu} Z(\mu, \nu) \ \mu > 0, \ \nu \geq 0, \tag{1.35}$$

where ν is a dispersion parameter and $Z(\mu, \nu) = \sum_{j=0}^{\infty} \frac{\mu^j}{(j!)^\nu}$ normalizes the distribution. Similar to the generalized Poisson pmf, when the dispersion parameter $\nu = 1$,

(1.33) reduces to the Poisson pmf, while $\nu > 1$ and $\nu < 1$ characterize overdispersion and underdispersion, respectively. The flexibility with the CMP distribution is that it can capture two other classic discrete distributions, namely the geometric distribution with success probability $(1 - \mu)$ when $\nu = 0$ and $\mu < 1$, and the Bernoulli distribution with success probability $\mu/(1 + \mu)$ when $\nu \to \infty$. [87] later proposed a generalized CMP distribution that generalizes both the CMP distribution and the negative binomial distribution. [88] first proposed a CMP regression model, where μ is related to a vector of covariates $\boldsymbol{x}$ using the log link function. Just like the CMP distribution generalizes several different discrete distributions, the CMP regression model generalizes both Poisson and logistic regression models.

[39] introduced a ZICMP regression model when excess zeros are present in a CMP regression setting. The authors further allowed the dispersion parameter to be modeled as a function of covariates via the log link. The probability of observing a zero from the degenerate state, π, is again allowed to be modeled as a function of covariates via a logit link. Just like the discussion of ZI count regression models in Section 1.2, the covariates used when modeling the parameters μ, ν, and π need not all be the same. [39] also developed the LRT for the presence of significant data dispersion, derived the Fisher information matrix for computing the estimated parameter standard errors, and conducted a broad simulation study comparing the ZICMP regression model fit to other standard ZI count regression fits. The model was demonstrated to provide a nearly similar fit (in terms of its log-likelihood) relative to the ZINB and ZIG regression fits, thus indicating the ZICMP regression's ability to characterize data dispersion. The authors have also made available their functions related to this work in the `R` package `COMPoissonReg` [89].

1.8 ZI Models for Clustered Data

Longitudinal or panel study designs can also result in longitudinal or clustered ZI count data. As noted in Feng and Zhu [90], ignoring the within-cluster correlation of longitudinal data will lead to loss of efficiency and incorrect inference on the regression coefficients. Most research in handling longitudinal ZI count data has been restricted to the ZIP regression setting. In particular, a marginal model and a conditional model for ZIP regression are two approaches commonly taken in the literature.

Hall and Zhang [26] framed the approach for finding the MLEs in marginal ZIP regression models by using generalized estimating equations (GEEs). Following their discussion, let $\boldsymbol{y}_i \in \mathbb{R}^{n_i}$ be a vector of responses for the i^{th} cluster, $i = 1, \ldots, M$. In

a marginal ZI count regression model, the random variable Y_{ij} associated with the observation y_{ij}, $j = 1, \ldots, n_i$, follows a ZI distribution as defined in Section 1.2, but where the count distribution must belong to the exponential dispersion family [91]. Let R_{ij} be the indicator variable that Y_{ij} came from the degenerate distribution at 0. Under independence, the complete data log-likelihood based on $\boldsymbol{y} = (\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_M^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{R} = (r_{11}, \ldots, r_{Mn_M})^{\mathrm{T}}$ separates as follows:

$$\ell_c(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi; \boldsymbol{y}, \boldsymbol{r}) = \ell_c(\boldsymbol{\pi}(\boldsymbol{\alpha}); \boldsymbol{y}, \boldsymbol{r}) + \ell_c(\boldsymbol{\mu}(\boldsymbol{\beta}), \phi; \boldsymbol{y}, \boldsymbol{r}),$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}) = (\mu_1(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta}))^{\mathrm{T}}$ and $\boldsymbol{\pi}(\boldsymbol{\alpha}) = (\pi_1(\boldsymbol{\alpha}), \ldots, \pi_n(\boldsymbol{\alpha}))^{\mathrm{T}}$ have been used to generically represent the conditional mean of both components, and we have replaced $\boldsymbol{\vartheta}$ in (1.2) with the univariate scale parameter ϕ as defined in the exponential dispersion family. Using an EM algorithm, at the t^{th} iteration we maximize

$$\begin{aligned} \mathcal{Q}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi | \boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \phi^{(t)}\right) &= \ell_c\left(\boldsymbol{\pi}(\boldsymbol{\alpha}); \boldsymbol{y}, \widehat{\boldsymbol{r}}^{(t)}\right) \\ &\quad + \ell_c\left(\boldsymbol{\mu}(\boldsymbol{\beta}), \phi; \boldsymbol{y}, \widehat{\boldsymbol{r}}^{(t)}\right), \end{aligned} \tag{1.36}$$

where $\widehat{\boldsymbol{r}}^{(t)}$ is an estimate of the posterior membership probabilities calculated in the E-step. The M-step requires maximizing $\mathcal{Q}$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and solving the following respective equations:

$$\sum_{i=1}^{M} \left\{ \frac{\partial \boldsymbol{\pi}_i(\boldsymbol{\alpha})^{\mathrm{T}}}{\partial \boldsymbol{\alpha}} \right\} \left[\left(\boldsymbol{A}_i(\boldsymbol{\pi}_i(\boldsymbol{\alpha}))\right)^{1/2} \boldsymbol{I}_{n_i} \left(\boldsymbol{A}_i(\boldsymbol{\pi}_i(\boldsymbol{\alpha}))\right)^{1/2} \right]^{-1} (\hat{\boldsymbol{r}}_i^{(t)} - \boldsymbol{\pi}_i(\boldsymbol{\alpha})) = \boldsymbol{0}, \tag{1.37}$$

$$\sum_{i=1}^{M} \left\{ \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \right\} \left[\left(\boldsymbol{B}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\right)^{1/2} \boldsymbol{I}_{n_i} \left(\boldsymbol{B}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\right)^{1/2} \right]^{-1} \boldsymbol{W}_i^{(t)} (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \boldsymbol{0}. \tag{1.38}$$

In the above, $\boldsymbol{I}_{n_i}$ is the $(n_i \times n_i)$ identity matrix, $\boldsymbol{A}_i(\boldsymbol{\pi}_i(\boldsymbol{\alpha})) = \mathrm{diag}(\pi_{i1}(\boldsymbol{\alpha})(1 - \pi_{i1}(\boldsymbol{\alpha})), \ldots, \pi_{in_i}(\boldsymbol{\alpha})(1 - \pi_{in_i}(\boldsymbol{\alpha})))$, $\boldsymbol{W}_i^{(t)} = \mathrm{diag}(1 - \widehat{r}_{i1}^{(t)}, \ldots, 1 - \widehat{r}_{in_i}^{(t)})$, and $\boldsymbol{B}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))$ is an $(n_i \times n_i)$ diagonal matrix with entries composed of the conditional variance; see [26] for how this last quantity is explicitly defined. In the above, the conditional mean μ and mixing proportion π from Section 1.2 have been vectorized and written explicitly as functions of the parameters to be estimated; i.e., $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ and $\boldsymbol{\pi}_i(\boldsymbol{\alpha})$. Then, the scale parameter ϕ needs to be estimated.

The formulas in (1.37) and (1.38) have the form of (weighted) GEEs with working correlation matrix equal to the identity matrix. Hall and Zhang [26] and Dobbie and

Welsh [92] explore substituting the working correlation structures in the marginal model approach with something other than the identity matrix, such as an exchangeable or AR(1) structure. To guard against correlation misspecification, Hall and Zhang [26] advocate using the GEE-1 approach of Liang et al. [93], which treats the first and second moment parameters orthogonally. Finally, Iddi and Molenberghs [94] extended the framework of Hall and Zhang [26] and presented a marginalized, ZI, overdispersed model for correlated data.

The basic framework of the conditional model approach is to use mixed effects models for $g(\mu)$ and $h(\pi)$. This approach was first considered in Hall [24] for ZIP and ZIB (zero-inflated binomial) regression with random intercepts, where the parameters were estimated using an EM algorithm. Wang et al. [95] obtained the penalized likelihood function by treating the random effects as unknown parameters, and then using residual maximum likelihood (REML) for estimation. [96] and [97] have also proposed mixed effects models to accommodate within-subject and between-subject heterogeneity in the presence of zero-inflation. [98] take a semiparametric approach to the model in [96] by relaxing the normality assumption of the random effects and leaving the corresponding distribution unspecified.

For the mixed model version of ZIP regression, it is typically assummed that $y_{ij}|a_i, b_i \sim \text{ZIP}(\pi_{ij}(\boldsymbol{\alpha}, a_i), \mu_{ij}(\boldsymbol{\beta}, b_i))$ where $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ represent the i$^{\text{th}}$ subject j^{th} observation within subject, respectively. Similar to before, $\mu_{ij}(\boldsymbol{\beta}, b_i) = \exp(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sigma_b b_i)$ and $\pi_{ij}(\boldsymbol{\alpha}, a_i) = \text{logit}^{-1}(\boldsymbol{w}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \sigma_a a_i)$. Moreover, it is typically assumed that the random intercepts $a_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and $b_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ for $i = 1, \ldots, n$. Typically, $a_i \perp b_i$, but some spatial models consider correlation of the two random intercepts; see [99]. Also, some authors do not include the random intercept in the zero-inflation state due the inability to establish heterogeneity in a latent process [46]. The most common method of estimation for mixed models is maximizing the marginal likelihood via Gauss-Quadrature. The likelihood function for the mixed model is

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_a^2, \sigma_b^2; \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) = &\sum_{i=1}^{n}\sum_{j=1}^{n_i} \log(f(y_{ij}|a_i, b_i; \pi_{ij}(\boldsymbol{\alpha}, a_i), \mu_{ij}(\boldsymbol{\beta}, b_i))) + \sum_{i=1}^{n} \log(\phi(a_i; \sigma_a^2)) \\ &+ \sum_{i=1}^{n} \log(\phi(b_i; \sigma_b^2)),\end{aligned} \tag{1.39}$$

where $f(y_{ij}|a_i, b_i; \pi_{ij}(\boldsymbol{\alpha}, a_i), \mu_{ij}(\boldsymbol{\beta}, b_i))$ is the ZIP mass function, and $\phi(\cdot; \sigma^2)$ is the

Gaussian density with variance σ^2. Now, the random intercepts are unobservable, and so we instead maximize the marginal likelihood

$$\ell_m(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_a^2, \sigma_b^2; \boldsymbol{y}) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \ell(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_a^2, \sigma_b^2; \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) d\boldsymbol{b} d\boldsymbol{a}, \tag{1.40}$$

which typically cannot be evaluated in closed form. Instead, we approximate the integral using Gauss-Hermite quadrature. Gauss-Hermite quadrature approximates integrals of $\int_{\mathbb{R}} f(x)e^{-x^2}dx$ by $\sum_{k=1}^{K} w_k f(x_k)$, where x_k are the nodes and w_k are the weights.

Maximization of (1.38) can be difficult due to the integration with respect to a_i and b_i [24]. To mitigate this, we can treat $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{r})$ as missing data, where $\boldsymbol{r}$ denotes the posterior memberships, and apply an EM algorithm with Gaussian quadrature to obtain the MLEs. Let $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_a^2, \sigma_b^2)$. The complete data likelihood is

$$\begin{aligned}
\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{a}, \boldsymbol{b}) &= \log f(\boldsymbol{a}; \sigma_a^2) + \log f(\boldsymbol{b}; \sigma_b^2) + \log f(\boldsymbol{y}, \boldsymbol{r} | \boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \phi(a_i; \sigma_a^2) + \sum_{i=1}^{n} \log \phi(b_i; \sigma_b^2) \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{n_i} \Bigg[\Big(r_{ij}(\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha} + a_i \sigma_a) - \log(1 + e^{\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha} + a_i \sigma_a}) \Big) \\
&+ (1 - r_{ij}) \Big(y_{ij}(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + b_i \sigma_b) - \exp(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + b_i \sigma_b) - \log(y_{ij}!) \Big) \Bigg].
\end{aligned} \tag{1.41}$$

The EM Algorithm is:

1. **E-Step** - Calculate

$$Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t)}) = \mathbb{E}(\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{a}, \boldsymbol{b}) | \boldsymbol{y}, \boldsymbol{\Theta}^{(t)}), \tag{1.42}$$

where the expectation is taken with respect to the distribution $(\boldsymbol{r}, \boldsymbol{a}, \boldsymbol{b}) | \boldsymbol{y}$ and the current estimates of the parameters $\boldsymbol{\Theta}^{(t)}$. Then, by law of iterated expectation,

$$Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t)}) = \mathbb{E}\Big[\mathbb{E}(\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{a}, \boldsymbol{b}) | \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Theta}^{(t)}) | \boldsymbol{y}, \boldsymbol{\Theta}^{(t)} \Big], \tag{1.43}$$

where the inner expectation is with respect to $\boldsymbol{r} | \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{y}$, and the outer expectation is taken with respect to $\boldsymbol{a}, \boldsymbol{b} | \boldsymbol{y}$. Since $\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{a}, \boldsymbol{b})$ is linear in $\boldsymbol{r}$, the inner expectation

becomes $\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{r}^{(t+1)}, \boldsymbol{a}, \boldsymbol{b})$, where

$$
\begin{aligned}
r_{ij}^{(t+1)} &= \mathbb{P}(R_{ij} = 1 | \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\Theta}^{(t)}) \\
&= \begin{cases} 0 & y_{ij} > 0 \\ \left[1 + \exp\Big(- (\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha}^{(t)} + \sigma_a^{(t)} a_i + e^{\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}^{(t)} + \sigma_b^{(t)} b_i}) \Big) \right]^{-1} & y_{ij} = 0. \end{cases}
\end{aligned} \tag{1.44}
$$

We'll write $r_{ij}^{(t+1)}(a_i, b_i)$ to denote that r_{ij} depends on (a_i, b_i).

Now, we need to compute the outer expectation with respect to $(\boldsymbol{a}, \boldsymbol{b})|\boldsymbol{y}$:

$$
\begin{aligned}
Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) &= \mathbb{E}(\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{r}^{(t)}, \boldsymbol{a}, \boldsymbol{b})) \\
&\propto \sum_{i=1}^{n} \sum_{j=1}^{n_i} \int_{\mathbb{R}} \int_{\mathbb{R}} \Bigg\{ \big[r_{ij}^{(t+1)}(a_i, b_i)(\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha} + \sigma_a a_i) - \log(1 + e^{\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha} + \sigma_a a_i}) \big] \\
&\quad + (1 - r_{ij}^{(t+1)}(a_i, b_i)) \times \big[y_{ij}(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \sigma_b b_i) - \exp(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \sigma_b b_i) \big] \Bigg\} \\
&\quad \times f(a_i, b_i | \boldsymbol{y}_i; \boldsymbol{\Theta}^{(t)}) da_i db_i \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n_i} \int_{\mathbb{R}} \int_{\mathbb{R}} \Bigg\{ \big[r_{ij}^{(t+1)}(a_i, b_i)(\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha} + \sigma_a a_i) - \log(1 + e^{\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha} + \sigma_a a_i}) \big] \\
&\quad + (1 - r_{ij}^{(t+1)}(a_i, b_i)) \times \big[y_{ij}(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \sigma_b b_i) - \exp(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \sigma_b b_i) \big] \Bigg\} \\
&\quad \times f(\boldsymbol{y}_i | a_i, b_i; \boldsymbol{\Theta}^{(t)}) \phi(a_i) \phi(b_i) da_i db_i \times \Big[\int_{\mathbb{R}} \int_{\mathbb{R}} f(\boldsymbol{y}_i | a_i, b_i; \boldsymbol{\Theta}^{(t)}) \phi(a_i) \phi(b_i) da_i db_i \Big]^{-1}.
\end{aligned} \tag{1.45}
$$

Here, $f(\boldsymbol{y}_i | a_i, b_i; \boldsymbol{\Theta}^{(t)} = \prod_{j=1}^{n_i} f(y_{ij} | a_i, b_i; \boldsymbol{\Theta}^{(t)})$, where $f(y_{ij} | a_i, b_i; \boldsymbol{\Theta}^{(t)})$ is the ZIP mass function. Then, employing Gauss-Hermite Quadrature, the E-Step becomes

$$
\begin{aligned}
Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) &\approx \sum_{i,j} \Bigg\{ \frac{\sum_{k=1}^{m_a} \sum_{l=1}^{m_b} \big[r_{ij}^{(t+1)}(a_k, b_l)(\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha} + \sigma_a a_k) - \log(1 + e^{\boldsymbol{w}_{ij}^{\mathrm{T}} \boldsymbol{\alpha} + \sigma_a a_k}) \big]}{\sum_{k=1}^{m_a} \sum_{l=1}^{m_b} f(\boldsymbol{y}_i | a_k, b_l; \boldsymbol{\Theta}^{(t)}) c_k q_l} \\
&\quad \times f(\boldsymbol{y}_i | a_k, b_l; \boldsymbol{\Theta}^{(t)}) c_k q_l \\
&\quad + \frac{\sum_{k=1}^{m_a} \sum_{k=1}^{m_b} \Big[(1 - r_{ij}^{(t+1)}(a_k, b_l)) \times \big[y_{ij}(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \sigma_b b_l) - \exp(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \sigma_b b_l) \big] \Big]}{\sum_{k=1}^{m_a} \sum_{l=1}^{m_b} f(\boldsymbol{y}_i | a_k, b_l; \boldsymbol{\Theta}^{(t)}) c_k q_l} \\
&\quad \times f(\boldsymbol{y}_i | a_k, b_l; \boldsymbol{\Theta}^{(t)}) c_k q_l \Bigg\},
\end{aligned} \tag{1.46}
$$

where (a_k, b_l) are the quadrature points with associated weights (c_k, q_l).
Finally, setting $v_{ijkl} = f(\boldsymbol{y}_i|a_k, b_l; \boldsymbol{\Theta}^{(t)})c_k q_l / g_i^{(t)}$, where $g_i^{(t)} = \sum_{k=1}^{m_a}\sum_{l=1}^{m_b} f(\boldsymbol{y}_i|a_k, b_l; \boldsymbol{\Theta}^{(t)})c_k q_l$, (1.44) becomes

$$\begin{aligned} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) &\approx \sum_{i,j,k,l} v_{ijkl}\left[r_{ij}^{(t+1)}(a_k, b_l)(\boldsymbol{w}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \sigma_a a_k) - \log(1 + e^{\boldsymbol{w}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \sigma_a a_k})\right] \\ &\quad + \sum_{i,j,k,l} v_{ijkl}(1 - r_{ij}^{(t+1)}(a_k, b_l)) \times \left[y_{ij}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sigma_b b_l) - \exp(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sigma_b b_l)\right] \\ &= \ell_c(\boldsymbol{\alpha}, \sigma_a; \boldsymbol{r}^{(t+1)}) + \ell_c(\boldsymbol{\beta}, \sigma_b; \boldsymbol{y}, \boldsymbol{r}^{(t+1)}). \end{aligned} \tag{1.47}$$

2. **M-Step** - Can be broken ito two steps:

   a) *M-Step* for $(\boldsymbol{\alpha}, \sigma_a)$: Maximize $\ell_c(\boldsymbol{\alpha}, \sigma_a; \boldsymbol{r}^{(t+1)})$ via logistic regression. Define the response vector for the (i, j) "observations" as

$$\boldsymbol{r}_{ij}^* = (r_{ij}(a_1, b_1), \ldots r_{ij}(a_1, b_{m_b}), r_{ij}(a_2, b_1), \ldots, r_{ij}(a_2, b_{m_b}), \ldots)^{\mathrm{T}}$$

   which is of length $m_a \times m_b$. Then, define the response vector for the i$^{\text{th}}$ subject as $\boldsymbol{r}_i^* = (\boldsymbol{r}_{i1}^{*\mathrm{T}}, \ldots, \boldsymbol{r}_{in_i}^{*\mathrm{T}})^{\mathrm{T}}$, and then the overall response vector as $\boldsymbol{r}^* = (\boldsymbol{r}_1^{*\mathrm{T}}, \ldots, \boldsymbol{r}_n^{*\mathrm{T}})^{\mathrm{T}}$, which is of length $N^* = \sum_{i=1}^{n}(n_i \times m_a \times m_b)$. Now define the parameter vector as $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}^{\mathrm{T}}, \sigma_a)^{\mathrm{T}}$. Note that here $\boldsymbol{r}^* = \boldsymbol{r}^{*(t+1)}$, where the $(t+1)$ superscript has been suppressed for convenience. Moreover, define the covariate vector for "observation" (i, j, k, l) as $\boldsymbol{w}_{ijkl} = (\boldsymbol{w}_{ij}^{\mathrm{T}}, a_k)^{\mathrm{T}}$. Note that the covariate vector $\boldsymbol{w}_{ijkl}$ is constant across the l subscript; i.e., $\boldsymbol{w}_{ijkl} = \boldsymbol{w}_{ijkl^*}$ for $l \neq l^*$. Define the matrix of explanatory variables for "observation" (i, j) by concatinating the $\boldsymbol{w}_{ijkl}$ for a fixed i and j; i.e.,

$$\boldsymbol{W}_{ij}^* = \left(\boldsymbol{w}_{ij11}, \ldots, \boldsymbol{w}_{ij1m_b}, \boldsymbol{w}_{ij21}, \ldots, \boldsymbol{w}_{ij2m_b}, \ldots\right)^{\mathrm{T}}.$$

   Then, define the $N^* \times (q+1)$ matrix of explanatory variables for all "observations" $\boldsymbol{W}^*$ as

$$\boldsymbol{W}^* = \begin{pmatrix} \boldsymbol{W}_{11}^* \\ \vdots \\ \boldsymbol{W}_{1n_1}^* \\ \boldsymbol{W}_{21}^* \\ \vdots \\ \boldsymbol{W}_{2m_2}^* \\ \vdots \end{pmatrix}.$$

Similarly, define the weight vector for "observation" (i, j) as

$$\boldsymbol{v}_{ij} = (v_{ij11}, \dots v_{ij1m_b}, v_{ij21}, \dots v_{ij2m_b}, \dots)^{\mathrm{T}},$$

and the weight vector for the i$^{\text{th}}$ "subject" as

$$\boldsymbol{v}_i = \left(\boldsymbol{v}_{i1}^{\mathrm{T}}, \dots, \boldsymbol{v}_{in_i}^{\mathrm{T}}\right)^{\mathrm{T}}.$$

Lastly, define the overall weight vector as $\boldsymbol{v} = (\boldsymbol{v}_1^{\mathrm{T}}, \dots, \boldsymbol{v}_n^{\mathrm{T}})^{\mathrm{T}}$. Now rewrite $\ell_c(\boldsymbol{\alpha}, \sigma_a; \boldsymbol{r}^{(t+1)})$ as

$$\ell_c(\boldsymbol{\alpha}, \sigma_a; \boldsymbol{r}^{(t+1)}) = \sum_{ijkl} v_{ijkl}\left[r_{ij}^{(t+1)}(a_k, b_l)(\boldsymbol{w}_{ijkl}^{\mathrm{T}}\boldsymbol{\alpha}^*) - \log(1 + e^{\boldsymbol{w}_{ijkl}^{\mathrm{T}}\boldsymbol{\alpha}^*})\right]. \quad (1.48)$$

Therefore, the maximization of $\ell_c(\boldsymbol{\alpha}, \sigma_a; \boldsymbol{r}^{(t+1)})$ can be accomplished via logistic regression of $\boldsymbol{r}^*$ on $\boldsymbol{W}^*$ with weights $\boldsymbol{v}$.

b) *M-Step* for $(\boldsymbol{\beta}, \sigma_b)$: Similar to the previous *M-Step*, define $\boldsymbol{y}_{ij}^* = y_{ij} \times \mathbf{1}_{m_a \times m_b}$, and $\boldsymbol{y}_i^* = (\boldsymbol{y}_{i1}^{*^{\mathrm{T}}}, \dots, \boldsymbol{y}_{in_i}^{*^{\mathrm{T}}})^{\mathrm{T}}$. Then, set $\boldsymbol{y}^* = (\boldsymbol{y}_i^{*^{\mathrm{T}}}, \dots, \boldsymbol{y}_n^{*^{\mathrm{T}}})^{\mathrm{T}}$. Define $\boldsymbol{x}_{ijkl} = (\boldsymbol{x}_{ij}^{\mathrm{T}}, b_l)^{\mathrm{T}}$, and then define the matrix of explanatory variables $\boldsymbol{X}^*$ in a similar manner to $\boldsymbol{W}^*$. Define the parameter vector as $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^{*^{\mathrm{T}}}, \sigma_b)^{\mathrm{T}}$. Finally, construct the weight vector $\boldsymbol{u}$ analougously to $\boldsymbol{v}$ with weights $u_{ijkl} = v_{ijkl}(1 - r_{ij}(a_k, b_l))$.
Then, rewrite $\ell_c(\boldsymbol{\beta}, \sigma_b; \boldsymbol{y}, \boldsymbol{r}^{(t+1)})$ as

$$\ell_c(\boldsymbol{\beta}, \sigma_b; \boldsymbol{y}, \boldsymbol{r}^{(t+1)}) = \sum_{ijkl} u_{ijkl} \times \left[y_{ij}(\boldsymbol{x}_{ijkl}^{\mathrm{T}}\boldsymbol{\beta}^*) - \exp(\boldsymbol{x}_{ijkl}^{\mathrm{T}}\boldsymbol{\beta}^*)\right]. \quad (1.49)$$

Therefore, the *M-Step* can be performed via Poisson regression of $\boldsymbol{y}^*$ on $\boldsymbol{X}^*$ with weight vector $\boldsymbol{u}$.

3. Iterate the *E-Step* and *M-Step* from $t = 0, 1, \dots$ until convergence.

For more complicated random effect distributions, such as the conditionally autoregressive distribution for spatial processes, authors resort to Bayesian techniques for estimation; see [46] and [99].

Count time series with extra zeros have also been explored in the literature. [95] was one of the first papers to consider this general setup, and developed a Markov ZIP regression model that allows for the frequency distribtuion to change according to the states of a two-state discrete time Markov chain with the transition probabilities associated with covariates through a logit link function. The model was then used for analyzing the daily number of phone calls on a fault report. A similar Markov

ZIP regression model was developed in [100], who employed a partial likelihood for conducting statistical inference. Both these models employ an random effect with an AR(1) covariance matrix. Furthermore, ZI integer-valued generalized autoregressive conditional heteroskedasticity models have been developed for the negative binomial and compound Poisson distribution in [101] and [102], respectively.

1.9 Zero-Inflation and Diagonal-Inflation in Multivariate Count Responses

Multivariate ZI models have been treated far less in the literature compared to their univariate counterparts, especially in the presence of covariates. The practical implications of multivariate ZI count regression models is that they foster descriptions of how a vector of correlated ZI count variables respond simultaneously to changes in measured covariates. Some applications of multivariate ZI regression models include development of a bivariate ZIP regression model for analyzing two types of occupational injuries (musculoskeletal and non-musculoskeletal) at a teaching hospital during different intervention trial time periods [95], a semiparametric bivariate ZIP regression model for analyzing two populations of fish (common carp and channel catfish) as a function of various environmental variables [103], and a bivariate ZINB regression model [95] and a bivariate ZIGP regressioon model [104] for analyzing healthcare utilization (doctor and non-doctor health professional visits) as a function of various socio-economic variables.

Consider a bivariate random yector $\boldsymbol{Y} = (Y_1, Y_2)^{\mathrm{T}}$, which has support on $\mathbb{N} \times \mathbb{N}$. The the most common type of inflation in multivariate count data is straightforward zero-inflation. In other words, there are excessive observed counts of $\boldsymbol{y} = (0, 0)^{\mathrm{T}}$. Another type of multivariate inflation is *diagonal-inflation*, where the counts $y_1 = y_2 = c$ for $c \in \mathbb{N}$ are observed at high frequencies. Diagonal inflation regression models have been used to model pre and post treatment studies, where the treatment may not have an effect on some patients for an unknown reason, and the number of draws in various sports games; see [105].

1.10 Related Models

There are various models available for handling other issues with zero counts in an observed dataset. Such models are often discussed along with ZI count regression models. We briefly highlight some of these models in this section.

In contrast to ZI count regression models, which are two-component mixture models, *hurdle regression models* are two-part models where it is assumed that the positive

counts are generated from a different process than the zero counts. Employing a similar notation to (1.2), a hurdle model combines a zero-truncated count distribution with a point mass at 0 as follows :

$$f(y;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\vartheta}) = \pi(\boldsymbol{\alpha})\mathrm{I}\{y=0\} + (1-\pi(\boldsymbol{\alpha}))\frac{p(y;\mu(\boldsymbol{\beta}),\boldsymbol{\vartheta})}{1-p(0;\mu(\boldsymbol{\beta}),\boldsymbol{\vartheta})}\mathrm{I}\{y>0\} \tag{1.50}$$

Notice, in contrast to the ZI regression model, zeros must come from the degenerate component. Hurdle models were first proposed by [2] to analyze survey data on beverage consumption. The `pscl` and `glmmTMB` packages in R, along with `PROC NLMIXED` in SAS, can be used to estimate hurdle regression models.

A large number of zeros can also be present in continuous data, but the probability of yielding a zero under a continuous distribution is 0. The setting can often be characterized with a *semi-continuous variable*, which has a portion of responses equal to a single value (commonly 0) and a continous, often right-skewed distribution, for the remaining values [106]. As noted in [26], "for independent semi-continuous data, there is little motivation for such as a ZI normal, because all observed zeros are unambiguous; they necessarily come from the degenerate distribution, rather than from the nondegenerate continuous distribution." Thus, the likelihood for such a model factors into terms for the zero and non-zero data, similar to the hurdle regression model. See [107] for a discussion of ZI gamma regression and ZI lognormal regression models, who also used those models to analyze data involving Parkinson's disease and driving capabilities.

A popular *semi-continuous distribution* used in insurance to model incurred loss is the *Compound Poisson-gamma distribution*, which is a special case of the more general family of *Tweedie Distributions*. Let N be the number of claims on an insurance policy, with $N \sim \text{Poisson}(\lambda)$, and let $Z_1,\ldots,Z_N \sim \text{Gamma}(\alpha,\gamma)$ be the total loss with the i$^{\text{th}}$ claim, where $i = 1,\ldots,N$. Assume $Z_i \perp Z_j|N$ for $i \neq j$. Set $Y = \sum_{i=1}^{N} Z_i$, which is the total incurred loss with a policy holder, so that $Y|N \sim \text{Gamma}(N\alpha,\gamma)$ if $N > 0$, and $Y|N = 0 \sim 0$. Then, we are interested in modeling the marginal distribution of Y

$$\begin{aligned} f(y;\lambda,\alpha,\gamma) &= \int_{\mathbb{R}} f_{Y|N}(y|n)\mathbb{P}(N=n)dN \\ &= e^{-\lambda} + \sum_{n=1}^{\infty}\left[(\Gamma(n\alpha)(\gamma)^{n\alpha})^{-1}y^{n\alpha-1}e^{-\frac{y}{\gamma}}\right]\left[(n!)^{-1}e^{-\lambda}\lambda^{n}\right] \\ &= e^{-\lambda} + e^{-\lambda-\gamma^{-1}y}y^{-1}\sum_{n=1}^{\infty}\frac{(\frac{\lambda y^{\alpha}}{\gamma^{\alpha}})^{n}}{\Gamma(n\alpha)\Gamma(n+1)}, \end{aligned} \tag{1.51}$$

which cannot be computed in closed form, however, numerical approximations utilizing series expansions and Fourier inversion can be employed; see [108].

As mentioned before, the *compound Poisson-gamma distribution* is a member of the *Tweedie distributions*. While other definitions exist, the Tweedie distributions can be defined as any random variable $X_+ = \sum_{i=1}^{N} X_i$ for which each X_i is random sample from an exponential dispersion family with the same canonical parameter and possibly different index parameters [91]. *Tweedie distributions* have the property that if $\mathbb{E}(X_+) = \mu$, then $\text{Var}(X_+) = \phi\mu^p$, where ϕ is a scale parameter and $p = 0$ or $1 < p \leq \infty$. This family contains many familiar distributions for varying index parameters p, such as the normal for $p = 0$, quasi-Poisson for $p = 1$, and gamma for $p = 2$. It can be shown that the *compound Poisson-gamma distribution* has the index parameter p such that $1 < p < 2$ [109]. For the random variable Y above, the Tweedie parameterization is

$$p = \frac{\alpha + 2}{\alpha + 1}, \quad \mu = \lambda\alpha\gamma, \quad \phi = \frac{\lambda^{1-p}(\alpha\gamma)^{2-p}}{2 - p}. \tag{1.52}$$

Furthermore, as p gets closer to 1, the distribution is closer to a Poisson distribution, whereas p closer to 2 shifts the distribution closer to a gamma distribution. Lastly, for all p such that $1 < p < 2$, the distribution has a positive point mass a zero, and is continuous for $X_+ > 0$.

For *Tweedie regression*, it is typically assumed that $\log(\mu) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$ for a vector of covariates $\boldsymbol{x}$. Frequently, it is also neccesary to model the dispersion parameter ϕ via $\log(\phi) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\xi}$ for a vector of covariates $\boldsymbol{w}$. Then, estimation of $(\boldsymbol{\xi}, \boldsymbol{\beta})$ is can be performed via optimization methods for generalized linear models such as IRLS or Newton-Raphson. Typically, estimation of p is done via grid search or by profile likelihood [110]. Discussion of tweedie regression for insurance data can be found in [109]. Software for the tweedie regression model include `PROC HPGENSELECT` in `SAS`, and the *glm* function in `R` with the `tweedie` package [111].

Related to the ZI regression models for semi-continuous data is the zero-one inflated beta (ZOIB) regression model, which was introduced by [112]. ZOIB regression can be used to model proportions with a high amount of observed zero and one proportions. For example, [113] took a Bayesian approach to estimate the parameters of a ZOIB regression model for modeling US county poverty rates, which yielded comparable results to the US Census Bureau's current small-area model for county poverty estimation. [19] developed a ZOIB regression model to analyze grid-cell data of a forest coverage ration as a function of two covariates. The `zoib` package [114] in

R can perform Bayesian estimation and inference for ZOIB regression models.

Morevover, tree-based methods for zero-inflated count data were explored in [115]. The authors proposed the ZIP likelihood as a purity measure within a node. The ZIP tree was then applied to the soldering data of AT&T, which was analyzed in [3].

The term *zero-altered models* (ZAP) is also found in the literature, which refers to either hurdle models or, more generally, any model that reflects some secondary behaviour of zero counts. Besides the models we discussed here, one could also have *zero-truncated models*, where the mixing proportion in the ZI model is allowed to be negative and, hence is no longer a true mixture distribution. More discussion on the differences between these different types of zero-altered models can be found in Chapter 11 of [22] and Chapter 17 of [65].

1.11 Example: Relationship Data

[116] presented data on $n = 387$ responses to a version of the Relational Pursuit-Pursuer Short Form (RP-PSF), which was used to study the unwanted pursuit behavior (UPB) of recently split couples. The form consisted of 28 questions about the pursuer's behavior — e.g., "Did the pursuer leave unwanted gifts?" — each measured on a five-point Likert scale (from 0 for *never* to 4 for *over five times*). The response UPB is a discrete summary index to these 28 questions, where higher scores indicate more perpetration. These data were analyzed using ZI count regression models by [117] and [39], where the latter noted the clear presence of overdispersion since the mean UPB is 2.284 and the corresponding variance is 23.302. The predictors of interest are the anxious attachment level (continuous) between the previous couple, and a binary indicator for education level (0 for *lower than a bachelor's degree* and 1 for *at least a bachelor's degree*).

Figure 1.4 is a histogram of the frequency of UPB counts. The frequency was truncated at 15 in order to focus on the majority of the data. There are nine UPB counts greater than 15, and the maximum observed count is 34. We overlaid the fits based on the Poisson, negative binomial, ZIP, and ZINB distributions. Clearly the Poisson and ZIP fits are not appropriate as they noticeably deviate from the general shape of the data. The negative binomial and ZINB fits, however, provide a noticeable improvement. These fits were all obtained without accounting for covariates, but the observations we made with the histogram suggest that using the negative binomial or ZINB distribution for the count regression distribution should be a reasonable choice.

We next performed the boundary LRTs discussed in Section 1.3. The test of zero-

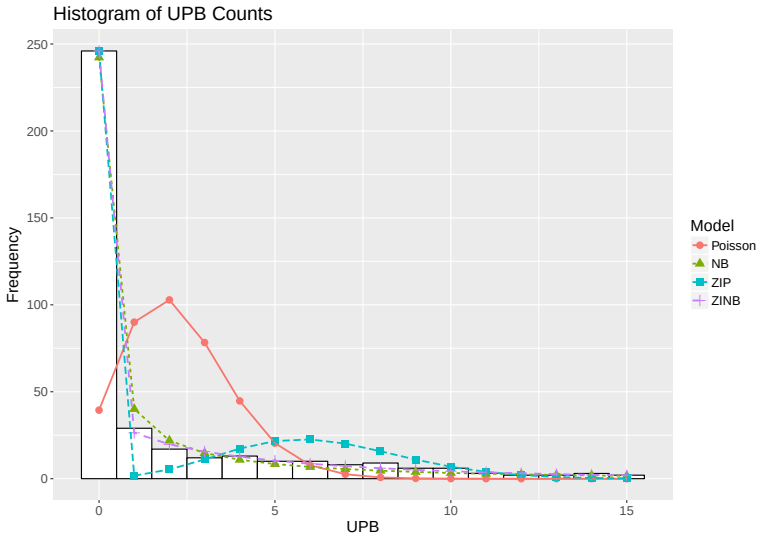


Figure 1.4: Histogram of the relationship data, truncated to show values of 15 or fewer for the UPB response. Fits for the four count distributions — Poisson, negative binomial, ZIP, and ZINB — are overlaid. A visually better fit can be seen with the estimated ZINB model.

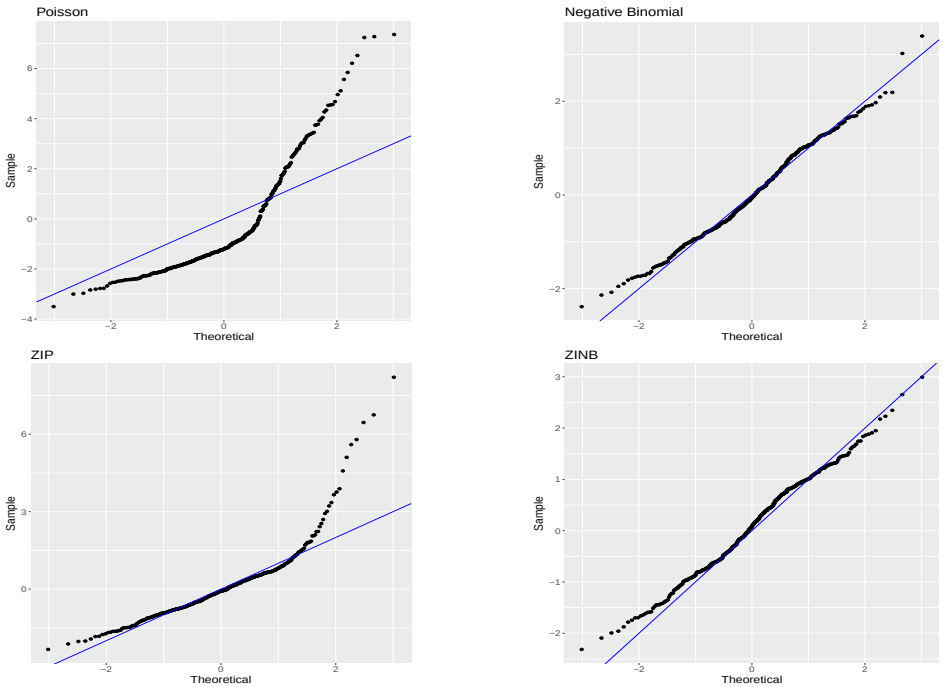


Figure 1.5: Q–Q plots of the randomized residuals for the fitted models: (a) Poisson regression, (b) negative binomial regression, (c) ZIP regression, and (d) ZINB regression. Better fits are indicated by the negative binomial regression and ZINB regression Q–Q plots.

Table 1.7: Adjusted-R^2 results when including education level, anxiety attachment level, or both in the model. Results for both the ZIP regression and ZINB regression models are reported.

Covariates	**ZIP Regression**	**ZINB Regression**
Education	0.523	0.016
Attachment	0.492	0.026
Education, Attachment	0.503	0.028

inflation for the Poisson regression and negative binomial regression settings, as well as the test for using a ZIP versus a ZINB regression model, all have highly significant results in favor of the alternative hypotheses; the largest p-value is 2.06×10^{-7}. Thus, these tests indicate the presence of zero-inflation and, more specifically, the use of the ZINB distribution. Table 1.7 gives the adjusted-R^2 values for the ZIP and ZINB regression models when including education level, anxiety attachment level, or both covariates in the respective model. These covariates were included in both the conditional mean model for the count distribution and the mixing proportion model for the zero inflation. For the ZIP regression models, the largest adjusted-R^2 is obtained for the model with only education level as a covariate. For the ZINB regression models, the largest adjusted-R^2 is obtained for the model with both covariates included. Since the boundary LRTs indicated the use of the ZINB distribution, we use the adjusted-R^2 results from this model and include both covariates in the model. As one final check of the fit, we calculated the randomized quantile residuals for the Poisson regression, negative binomial regression, ZIP regression, and ZINB regression models, where both covariates are included. The quantile-quantile (Q–Q) plots for these four estimated models are given in Figure 1.5. Clearly, better fits are obtained using negative binomial regression or ZINB regression. In fact, the ZINB regression model provides a slightly better fit for those values in the right-hand tail of the distribution.

Lastly, a table of AIC and BIC values can be seen in Table 1.8. Again, we see evidence that the ZINB provides the best fit, although the negative binomial provides a quality fit as well.

Table 1.8: AIC and BIC Values for Couple Data

Model	**AIC**	**BIC**	**ΔAIC**	**ΔBIC**
Poisson	2782.390	2794.266	1516.108	1500.275
NB	1285.919	1301.752	19.637	7.761
ZIP	1616.901	1640.652	350.619	346.661
ZINB	1266.282	1293.991	*	*

1.12 Appendix

ECM Algorithm for ZINB

Suppose we observe a sample of size n, $(y_1, \boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{w}_1^{\mathrm{T}}), \ldots, (y_n, \boldsymbol{x}_n^{\mathrm{T}}, \boldsymbol{w}_n^{\mathrm{T}})$, where the y_i are count responses, and $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ are (possibly uncoupled) p and q dimensional vectors of covariates, respectively. Following similar notation as used in Section 1.2 of the main text, the conditional distribution of the counts given the covariates for ZINB regression is

$$f(y_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \pi_i(\boldsymbol{\alpha})\mathrm{I}\{y_i = 0\} + (1 - \pi_i(\boldsymbol{\alpha}))p(y_i; \mu_i(\boldsymbol{\beta}), \theta) \tag{1.53}$$

where

$$p(y_i; \boldsymbol{\beta}, \theta) = \frac{\Gamma(\theta + y_i)}{y_i!\Gamma(\theta)} \left(\frac{\mu_i(\boldsymbol{\beta})}{\theta + \mu_i(\boldsymbol{\beta})}\right)^{y_i} \left(\frac{\theta}{\theta + \mu_i(\boldsymbol{\beta})}\right)^{\theta}, \ \mu_i(\boldsymbol{\beta}), \theta > 0 \ \forall i. \tag{1.54}$$

Taking the traditional GLM approach for modeling $\mu_i(\boldsymbol{\beta})$ and $\pi_i(\boldsymbol{\alpha})$, we assume

$$\log(\mu_i(\boldsymbol{\beta})) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \quad \text{and} \quad \mathrm{logit}(\pi_i(\boldsymbol{\alpha})) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha}. \tag{1.55}$$

Incorporating (1.55) into the loglikelihood based on (1.53), the observed data loglikelihood for the ZINB regression model is as follows:

$$\begin{aligned}
\ell_o(\boldsymbol{\beta}, \boldsymbol{\alpha}, \theta; \boldsymbol{y}) = &\sum_{y_i=0} \log\left(e^{\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha}} + (1 + \theta^{-1}e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}})^{-\theta}\right) \\
&- \sum_{y_i>0} \left[y_i \log\left(1 + \theta e^{-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right) + \theta \log(1 + \theta^{-1}e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}})\right] \\
&- \sum_{i=1}^{n} \log(1 + e^{\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha}}) + \sum_{y_i>0} \log\left(\frac{\Gamma(\theta + y_i)}{y_i!\Gamma(\theta)}\right).
\end{aligned}$$

Now suppose we knew which state the zeros came from; i.e., suppose we could observe

$$R_i = \begin{cases} 1, & \text{if } Y_i \text{ is from degenerate state;} \\ 0, & \text{if } Y_i \text{ is from the negative binomial state.} \end{cases}$$

In other words, R_i is a Bernoulli random variable with rate of success $\text{logit}^{-1}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha})$. Then, the complete data log-likelihood for the ZINB regression model is as follows:

$$\begin{aligned} \ell_c(\boldsymbol{\beta}, \boldsymbol{\alpha}, \theta; \boldsymbol{y}, \boldsymbol{r}) &= \sum_{i=1}^{n} \log(f(r_i|\boldsymbol{\alpha}) f(y_i|r_i, \boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} \left[r_i \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha} - \log(1 + e^{\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha}}) \right] \\ &\quad - \sum_{i=1}^{n} (1 - r_i) \left[y_i \log\left(1 + \theta e^{-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right) + \theta \log(1 + \theta^{-1} e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}) \right] \\ &\quad + \sum_{i=1}^{n} (1 - r_i) \log\left(\frac{\Gamma(\theta + y_i)}{y_i! \Gamma(\theta)}\right) \\ &= \ell_c(\boldsymbol{\alpha}; \boldsymbol{r}) + \ell_c(\boldsymbol{\beta}, \theta; \boldsymbol{y}, \boldsymbol{r}) \end{aligned}$$

where the $\boldsymbol{r} = (r_1, \ldots, r_n)^{\mathrm{T}}$ are the hypothetically observed indicators. Thus, we can maximize $\ell_c(\boldsymbol{\alpha}; \boldsymbol{r})$ and $\ell_c(\boldsymbol{\beta}, \theta; \boldsymbol{y}, \boldsymbol{r})$ separately in an EM [118] framework. However, note that maximizing $\ell_c(\boldsymbol{\beta}, \theta; \boldsymbol{y}, \boldsymbol{r})$ is actually difficult since the negative binomial distribution with unknown dispersion parameter does not belong to the exponential family. But if we consider two separate conditional maximization steps for maximizing $\ell_c(\boldsymbol{\beta}, \theta; \boldsymbol{y}, \boldsymbol{r})$, then we can more easily compute the MLEs for the ZINB regression model using an ECM algorithm [119].

ECM Algorithm

- **E-Step**: Let $\mathcal{Q}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)}) = \mathbb{E}[\ell_c(\boldsymbol{\beta}, \boldsymbol{\alpha}, \theta; \boldsymbol{y}, \boldsymbol{r}); \boldsymbol{\Theta}^{(t)}]$, where $\boldsymbol{\Theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \theta)^{\mathrm{T}}$ is our parameter of interest and $\boldsymbol{\Theta}^{(t)}$ is the corresponding estimate at iteration $t = 0, 1, \ldots$. The value of $\boldsymbol{\Theta}^{(0)}$ corresponds to user-supplied starting values. Estimate the probability of belonging to the perfect state by computing the posterior membership probabilities at the t^{th} iteration as follows:

$$r_i^{(t+1)} = \mathbb{P}\left(R_i = 1 | y_i, \boldsymbol{\Theta}^{(t)}\right) = \begin{cases} \left(1 + e^{\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha}^{(t)}} \left(1 + \frac{e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}}}{\theta^{(t)}}\right)^{-\theta^{(t)}}\right)^{-1}, & \text{if } y_i = 0; \\ 0, & \text{if } y_i = 1, 2, \ldots. \end{cases}$$

- **CM-Steps**:

    1. Calculate $\boldsymbol{\alpha}^{(t+1)}$ by maximizing $\mathcal{Q}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)})$ with $\boldsymbol{\beta}$ and θ fixed at $\boldsymbol{\beta}^{(t)}$ and $\theta^{(t)}$, respectively. In other words, this maximizes $\ell_c(\boldsymbol{\alpha}; \boldsymbol{y}, \boldsymbol{r}^{(t+1)})$ with respect to $\boldsymbol{\alpha}$, which can be solved by performing an unweighted binomial logistic regression of $r_1^{(t)}, \ldots, r_n^{(t)}$ on $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ [24].
    2. Calculate $\boldsymbol{\beta}^{(t+1)}$ by maximizing $\mathcal{Q}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)})$ with $\boldsymbol{\alpha}$ and θ fixed at $\boldsymbol{\alpha}^{(t+1)}$ and $\theta^{(t)}$, respectively. In other words, this conditionally maximizes $\ell_c(\boldsymbol{\beta}, \theta; \boldsymbol{y}, \boldsymbol{r}^{(t)})$ with respect to $\boldsymbol{\beta}$ by using a fixed value of θ, namely $\theta^{(t)}$. This is equivalent to estimating a weighted negative binomial regression model at a fixed dispersion parameter, which is a setting that puts the model in an exponential family. Thus, to calculate $\boldsymbol{\beta}^{(t+1)}$, we can employ IRLS as is typically done when estimating parameters in GLMs.
    3. Calculate $\theta^{(t+1)}$ by maximizing $\mathcal{Q}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)})$ with $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ fixed at $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\alpha}^{(t+1)}$, respectively. In other words, this conditionally maximizes $\ell_c(\boldsymbol{\beta}, \theta; \boldsymbol{y}, \boldsymbol{r}^{(t+1)})$ with respect to θ by using a fixed value of $\boldsymbol{\beta}$, namely $\boldsymbol{\beta}^{(t+1)}$. One could use, for example, a fixed point algorithm to calculate $\theta^{(t+1)}$ [120].

- Iterate between the E-step and the CM-steps until a convergence criterion is reached, such as $\ell_o(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \theta^{(t+1)}; \boldsymbol{y}) - \ell_o(\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \theta^{(t)}; \boldsymbol{y}) < \epsilon$ for some small $\epsilon > 0$.

Timing Results

In this section, we present the results from the simple time comparison study discussed in the main text. We generated a single dataset from a ZIP regression model and a ZINB regression model for each of the sizes $n \in \{10^k : k = 2, \ldots, 6\}$. For the ZIP regression model, the data were generated using $\boldsymbol{\beta} = (-2.0, 3.0)^{\mathrm{T}}$ and $\boldsymbol{\alpha} = (1.0, -1.7)^{\mathrm{T}}$. For the ZINB regression model, the data were generated using $\boldsymbol{\beta} = (3.0, 1.2)^{\mathrm{T}}$, $\boldsymbol{\alpha} = (-0.5, 0.5)^{\mathrm{T}}$, and $\theta = 6$. The results for estimating the ZIP regression model are given in Table 1.9. The results for estimating the ZINB regression model are given in Table 1.10. Estimating each model is almost always most efficient using `PROC GENMOD`.

Table 1.9: Timing results for estimating each dataset of size n from the generated ZIP regression model.

n	`zeroinfl`	`vglm`	`glmmTMB`	`GENMOD`	`NLMIXED`	`COUNTREG`
10^2	0.03	0.05	0.33	0.04	0.07	0.06
10^3	0.09	0.07	0.21	0.05	0.10	0.02
10^4	1.43	1.15	1.09	0.09	0.62	0.60
10^5	13.36	13.19	10.14	0.57	7.37	2.57
10^6	152.54	96.18	126.91	5.29	28.01	29.14

Table 1.10: Timing results for estimating each dataset of size n from the generated ZINB regression model.

n	`zeroinfl`	`vglm`	`glmmTMB`	`GENMOD`	`NLMIXED`	`COUNTREG`
10^2	0.02	0.31	0.30	0.16	0.55	0.25
10^3	0.19	2.45	0.38	0.15	0.22	0.18
10^4	2.78	22.90	3.13	0.20	0.55	0.75
10^5	23.73	225.36	38.71	1.02	3.25	6.18
10^6	227.80	2416.67	399.230	8.76	33.68	60.58

Insurance Analysis Code

Posted below is JAGS code for the insurance analysis. The JAGS model is fit using the "zeros trick". To describe the "zeros trick", let $\boldsymbol{Z}^* = \mathbf{0}_n$ be a "fake" response vectors, where each Z_i is an observed zero from a Poisson distribution with mean $\exp(\ell_i(\Theta) + K)$. Here, $\ell_i(\cdot)$ is the log-likelihood for the i$^{\text{th}}$ observation, and K is a sufficiently large constant, say 10000. K makes sure that $\ell_i(\Theta) + K > 0$ for all $i = 1, \ldots, n$. Then, the log-likelihood of the data is

$$\ell(\Theta) = \log(\prod_{i=1}^{n} P(Z_i = 0)) = \sum_{i=1}^{n} \big[\log(\ell_i(\Theta))\big] + n\log(K) \tag{1.56}$$

Thus, we can construct any arbitrary likelihood by employing the zeros trick. This trick, or the "ones trick", is necessary when construncting a likelihood for a distribution that is not already built into JAGS.

```
### JAGS Code - Save as .bug extension to run ###
## Vector of zeros for zeros trick
data{
K <- 10000
```

```
for(k in 1:N){
zeros[k] <- 0
}
}
## Constructing Model
model{
for(i in 1:N){
        ##Mixing Proportions
p[i] <- max(0.001,min(0.999,q[i]))
logit(q[i]) <- alpha[1] + alpha[2]*veh_value[i]
        ##Poisson Mean
log(mu[i]) <- log_exposure[i] + beta[1] + beta[2]*veh_value[i]+
      beta[3]*LargeVan_ind[i] + beta[4]*TwoSeat_ind[i]+
                      beta[5]*Convt_ind[i] +
      beta[6]*Bus_ind[i]+beta[7]*UTE_ind[i] +
      beta[8]*areaD[i] + beta[9]*agecat2[i] + beta[10]*agecat3[i] +
      beta[11]*agecat4[i] + beta[12]*agecat5[i] + beta[13]*agecat6[i]
        ##indicator for if y = 0 versus y >0
z[i] <- step(y[i] - 1)
        ## Likelihood Function
ll[i] <- (1-z[i])*log(p[i] + (1-p[i])*exp(-mu[i])) +
 z[i]*(log(1-p[i]) + y[i]*log(mu[i]) - mu[i] - loggam(y[i]+1))
## Likelihood construction for zeros trick (add large constant)
phi[i] <- -ll[i] + K
        ## Fake zeros are poisson with mean -ll[i]+K
zeros[i] ~ dpois(phi[i])
}
        ## Assigning normal priors with mean zero and
        ##precision .1 for regression coeff
for(j in 1:13){
beta[j] ~ dnorm(0,.1)
}
for(j in 1:2){
alpha[j] ~ dnorm(0,.1)
}
}
```

Furthermore, below is the R-Script using the RJAGS package [77], which allows JAGS scripts to be executed in the R environment.

```
################## File Information #########################
#### Purpose- Analyze Car Insurance Data From Australia ####

car <- read.csv("car.csv")
############## EDA ####################

car$veh_body <- as.character(car$veh_body)
## Setting up vehicle body groups
car <- transform.data.frame(car,
vh_body_group = ifelse(veh_body == "MCARA" | veh_body =="PANVN"
 ,"MCARA,PANVN",
ifelse(veh_body=="RDSTR" | veh_body=="COUPE","RDSTR,COUPE",
ifelse(veh_body=="BUS","BUS",
ifelse(veh_body=="CONVT"|veh_body=="HDTOP","CONVT",
ifelse(veh_body=="UTE","UTE","General"))))))

library(pscl)
car$agecat <- as.factor(car$agecat)
car$aread_ind <- ifelse(car$area == "D",1,0)
car$veh_age <- as.factor(car$veh_age)

########### Model Fitting ###########
##### Bayes #####
library(rjags)
## JAGS data should be input as a list
jags.data <- list('y'=car$numclaims,'veh_value'=car$veh_value,
'log_exposure'=log(car$exposure),
'LargeVan_ind'=as.numeric(car$veh_body=="MCARA"| car$veh_body=="PANVN"),
'TwoSeat_ind' = as.numeric(car$veh_body=="RDSTR" | car$veh_body=="COUPE"),
'Convt_ind'= as.numeric(car$veh_body=="CONVT"|car$veh_body=="HDTOP"),
'UTE_ind'= as.numeric(car$veh_body =="UTE"),
'Bus_ind'= as.numeric(car$veh_body =="BUS"),
'areaD'=(car$area=="D"),
```

```
'agecat2'=(car$agecat==2),'agecat3'=(car$agecat==3),
'agecat4'=(car$agecat==4),'agecat5'=(car$agecat==5),
'agecat6'=(car$agecat==6),'N'=nrow(car))
## Execture insurance_bayes.bug script with jags.data data set
jags <- jags.model("insurance_bayes.bug", data = jags.data)
## Obtain 5000 posterior samples
samples <- coda.samples(jags,
variable.names = c("alpha","beta"),n.iter = 5000)
```

Chapter 2 Semiparametric Extension to ZIP Regression via Local Likelihood

2.1 Introduction

In Chapter 1, ZI regression models were introduced, and discussed from a parametric perspective. ZI regression models are useful for studying the relationship between a discrete response variable with excessive zeros and predictor variables of interest. Recall that the ZI regression model can be written as the mixture model

$$Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w}\sim\pi(\boldsymbol{\alpha})\mathrm{I}\{y=0\}+(1-\pi(\boldsymbol{\alpha}))p(y;\mu(\boldsymbol{\beta}),\boldsymbol{\vartheta},\boldsymbol{x}), \tag{2.1}$$

where $\pi(\boldsymbol{\alpha})=h^{-1}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha})$ and $\mu(\boldsymbol{\beta})=g^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$ for suitably chosen link functions h and g. However, assuming $h^{-1}(\cdot)=\mathrm{logit}^{-1}(\cdot)$, the assumption of globally (logit) linear zero-inflation probabilities might be too strong, and thus, a more flexible model is desired. In this chapter, we propose relaxing the parametric assumption of the mixing proportions, and specify $\pi(\boldsymbol{w})$ as a smooth function of continuous covariates, where typically $\boldsymbol{w}\in\mathbb{R}^1$. The assumption of the parametric mean of the count component, $\mu(\boldsymbol{\beta})=g^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$, is still kept. Our semiparametric approach takes inspiration from the semiparametric mixtures-of-regressions literature, such as the work seen in [121] and [122]. For the rest of this chapter, we focus on the ZIP regression model, although one-parameter exponential families can be treated similarly. Before we formally introduce the semiparametric model and estimation, we first review the literature of semiparametric mixtures-of-regressions and semiparametric ZI regression.

Literature Review of Mixtures of Regressions

Mixtures-of-regression models, or "switching regression", were first developed in the econometrics literature by [123]. The authors discussed estimation of the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1+\epsilon_{1i} \ \text{ with probability } p \\ y_i &= \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2+\epsilon_{2i} \ \text{ with probability } 1-p, \end{aligned} \tag{2.2}$$

where $\epsilon_{1i}\overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma_1^2)$ and $\epsilon_{2i}\overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma_2^2)$ for $i=1,\ldots,n$. [123] discusses maximum likelihood estimation of the model, and applies the model to predict the number of housing starts by several economic variables. In 2.2, the switching probabilities p do

not depend on covariates.

The extension of allowing p to depend on covariates was developed in the machine learning literature by [124], which is called the *hierarchical mixtures of experts* (HME). Overall, the HME model is similar to the *Classification and Regression Trees* (CART) algorithm developed by [125], except the node splits are soft probabilistic splits as opposed to the the hard splits in CART. Also, within a node, the HME model fits a linear regression for prediction, as opposed to a constant (usually the mean) in CART. [124] develop an EM Algorithm to estimate the HME model.

Recently, semiparametric mixtures of regression models have been receiving increasing attention in the literature. [121] proposed a mixture-of-regressions model

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\phi}) = \sum_{j=1}^{m} \lambda_j(\boldsymbol{x}_i)\phi(\boldsymbol{y}_i;\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j,\sigma_j^2), \tag{2.3}$$

where the mixture components $\phi(\cdot;\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j,\sigma_j^2)$ are the Gaussian density with mean $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j$ and variance σ_j^2. Here, $\boldsymbol{\phi} = (\lambda_1(\cdot),\ldots,\lambda_m(\cdot),\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_m,\sigma_1^2,\ldots,\sigma_m^2)$. In their model, the mixing proportions $\lambda_j(\boldsymbol{x}_i)$ are assumed to be smooth function of covariates, which are then estimated via kernel regression. The authors develop strategies for choosing the number of mixture components m, and propose an "EM-like" algorithm for model estimation that alternates between local estimation of $\lambda_j(\cdot)$ and global estimation of each $\boldsymbol{\beta}_j$. The term "EM-like" will be further defined in Section 2. They then apply their semiparametric mixtures-of-regressions model to study how gross national product (GNP) per capita varies with the estimated carbon dioxide (CO_2) emissions per capita for a group of 28 nations. Thier model can be fit using the *mixtools* package in `R` [126].

[122] further developed on [121] by proposing a *one-step backfitting* algorithm for estimating 2.3, along with an "EM-like" for each of the back-fitting steps. The authors established novel asymptotic theory, which includes the asympotic normality of each $\widehat{\boldsymbol{\beta}}_j$, $\widehat{\sigma_j^2}$, and $\widehat{\lambda}_j(w)$, at a fixed point $w \in \mathbb{R}$. Also, the authors establish what they call the *asymptotic ascent property*, which will be further defined in Section 2. The asymptotic ascent property is analogous to the ascent property in classical EM algorithms. The *generalized likelihood ratio test* was developed by [127] to test the hypothesis

$$\begin{aligned} &\mathrm{H}_0 : \lambda_j(z) = \lambda_j \text{ for all } j = 1,\ldots,m \\ &\mathrm{H}_1 : \lambda_j(z) \neq \lambda_j \text{ for some } j = 1,\ldots,m. \end{aligned} \tag{2.4}$$

In other words, we are testing whether the mixing proportions depend on the covariates.

In similar work to [122], [128] developed a ZI binomial regression model to study how rainfall varies across time in Edmonton, Canada. Also, similar to [122], [129] developed a mixture of regressions model where the mixing proportions, mean functions for the components, as well as the variance functions for the components are modeled non-parametrically. The authors then apply their methodology to US housing price index data.

Literature Review of Semiparametric ZI Regression

Among the first semiparametric ZI regression model was the partially linear ZIP regression model developed by [97]. The authors developed the partially linear model where it is assumed that

$$\begin{aligned} \log(\mu_i) &= \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + g(T) \\ \operatorname{logit}(\pi_i) &= \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha}, \end{aligned} \tag{2.5}$$

T is an observed continuous predictor, and $g(\cdot)$ is an unknown smooth function. Denoting the parameter space by $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, g)$, and noting that $\boldsymbol{\Theta}$ is infinite dimensional. [97] make inference about $\boldsymbol{\Theta}$ via the sieve method. The key idea behind the sieve method is to "approximate the infinite dimensional parameter space $\boldsymbol{\Theta}$ by a sequence of finite dimensional parameter spaces $\boldsymbol{\Theta}_n$, where $\boldsymbol{\Theta}_n$ has larger dimension as $n \to \infty$." They then perform maximum likelihood estimation on $\boldsymbol{\Theta}_n$ instead of $\boldsymbol{\Theta}$.

More rigorously, suppose $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ lie in bounded subsets $A_1 \subset \mathbb{R}^p$ and $A_2 \subset \mathbb{R}^q$, respectively. For simplicity, suppose $T \in [0,1]$. Define the set

$$B = \{g \in C^r[0,1] : -\infty < m_0 \leq g(t) \leq M_0 < \infty, \forall t \in [0,1]\}, \tag{2.6}$$

where $C^r[0,1]$ denotes the class of r-order continuously differentiable functions on $[0,1]$. Then, $\boldsymbol{\Theta} = A_1 \times A_2 \times B$. The authors then approximate g by the B-spline basis

$$G_m(t; \boldsymbol{b}) = \sum_{j=1}^{m} \Big(\frac{b_j - b_{j-1}}{t_j - t_{j-1}} t - \frac{b_j t_{j-1} - b_{j-1} t_j}{t_j - t_{j-1}} \Big) \mathrm{I}\{t_{j-1} \leq t < t_j\}, \tag{2.7}$$

where $(t_0, \ldots, t_m)$ are the knots and $\boldsymbol{b} = (b_0, \ldots, b_m)^{\mathrm{T}}$ are the vector of coefficents for G_m. Here m is an integer with $m = O(n^k)$ for some $0 < k < 1$. Then, define

$$B_n = \{G_m(t; \boldsymbol{b}) : m_0 \leq b_j \leq M_0, j = 1, \ldots, m\}. \tag{2.8}$$

Let $\boldsymbol{\Theta}_n = A_1 \times A_2 \times B_n$ be the sieve space for $\boldsymbol{\Theta}$. For $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, g)^{\mathrm{T}} \in \boldsymbol{\Theta}$, select $\boldsymbol{b}^* = (g(t_0), \ldots, g(t_m))^{\mathrm{T}}$, and define the estimate of g as $g_n(\cdot) = G_m(\cdot; \boldsymbol{b}^*)$. Then, the

sieve MLE is defined as

$$\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\alpha}}_n, \widehat{g}_n)^{\mathrm{T}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n}{\operatorname{argsup}} \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}; y_i), \tag{2.9}$$

where $\ell_i(\boldsymbol{\theta}; y_i)$ is the ZIP likelihood for the i$^{\text{th}}$ observation. For the asymptotic properties, see [97]. For more details on sieve estimation, see [130] and [131]. [132] extended the sieve MLE work of [97] by assuming the zero-inflation probabilities $\text{logit}(\pi_i) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha} + h(T)$, where h is a smooth function of T.

Another novel work in the semiparametric ZI regression literature is the partially-constrained GAM developed by [133]. Similar to the ZIP(τ) model developed in [3], [133] allows for the effects of a predictor in the mean of the count component to be proportional to the effect in the zero-inflation probability. This allows us, for example, to answer the question of "does the temperature have similar influences in affecting the presence/absence of the speices and the local biomass, given the species is present in the location?". In more detail, assume $Y_i|\boldsymbol{T}_i, \boldsymbol{U}_i, \boldsymbol{V}_i, \boldsymbol{W}_i$ follows the ZI regression model for $i = 1, \ldots, n$, where the count component belongs in the one-parameter exponential family with mean μ_i. Here, $\boldsymbol{T}_i$, $\boldsymbol{U}_i$, $\boldsymbol{V}_i$, and $\boldsymbol{W}_i$ are vectors of covariates of length m_1, m_2, m_3, and m_4, respectively. Assume that the mean of the count component can be written as

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{m_1} s_j(\boldsymbol{t}_j) + \sum_{k=1}^{m_2} h_k(\boldsymbol{u}_k) + \sum_{l=1}^{m_3} \eta_l(\boldsymbol{v}_l), \tag{2.10}$$

where β_0 is an intercept, s_j, h_k, and η_l are non-parametric smooth functions. For identifiablity, we assume that each function has expectation zero. Similarly, suppose the zero-inflation probability can be written as

$$\text{logit}(\pi_i) = \alpha_0 + \delta_1 \sum_{j_1\in\mathcal{J}_1} s_{j_1}(\boldsymbol{t}_{j_1}) + \cdots + \delta_{m_1} \sum_{j_{m_1}\in\mathcal{J}_{m_1}} s_{j_{m_1}}(\boldsymbol{t}_{j_{m_1}}) + \sum_{k=1}^{m_2} h_k^*(\boldsymbol{u}_k) + \sum_{s=1}^{m_4} \xi_s(\boldsymbol{w}_s), \tag{2.11}$$

where $\mathcal{J}_1, \ldots, \mathcal{J}_{m_1}$ are subsets of the indices $\{1, \ldots, m_1\}$ with $\bigcup_{k=1}^{m} \mathcal{J}_k = \{1, \ldots m_1\}$. So, the covariate vectors of $\boldsymbol{v}_l$ and $\boldsymbol{w}_s$ are employed only in the count component and degenerate state, respectively. Moreover, the covariate vector $\boldsymbol{u}_j$ are used in both states, but the functions h_j and h_j^* are unrelated. Lastly, the covariate vector $\boldsymbol{t}_j$ is utilized in both states, but the effect in the zero-inflation state is proportional to that of the count component, where δ_j is the proportionally parameter. Model estimation is performed via penalized likelihood and an EM algorithm, and this model is used

to study jellyfish abundance by temperature, depth, and spatial-temporal variables. The `R` package `COZIGAM` [134] can be used to fit the constrained GAM model.

Lastly, [90] developed a semiparametric ZIP model for longitudinal data. Similar to the notation in Section 1.8, suppose we observe n indpendent subjects $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^{\mathrm{T}}$, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$. Then, assume that

$$Y_{ij}|b_i \sim \mathrm{ZIP}(\mu_{ij}(\boldsymbol{\beta}; b_i), \pi_{ij}(\boldsymbol{\alpha}))$$

and

$$\begin{aligned} \log(\mu_{ij}) &= \boldsymbol{x}_{ij}^{\mathrm{T}} + b_i + f(t_{ij}) \\ \mathrm{logit}(\pi_{ij}) &= \boldsymbol{w}_{ij}^{\mathrm{T}}\boldsymbol{\alpha}, \end{aligned}$$

where $f(\cdot)$ is a smooth function of a continuous covariate T, and $b_1, \ldots, b_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Here, f is approximated by splines with a q-order truncated power function as the basis; see [135] for more details. The authors then develop a penalized likelihood function with a *Monte-Carlo expectation-maximization* (MC-EM) algorithm for optimization. The methodology is then applied to a data set from a pharmaceutical company to monitor the number of side effect episodes on a patient.

Overview of Chapter

The rest of this chapter is organized as follows. Section 2.2 discusses estimation of our novel semiparametric ZIP regression model along with asymptotic properties of the estimators. Section 2.3 presents a simulation study for our model, along with a discussion of the Generalized Likelihood Ratio test. Section 2.4 shows an application of the semiparametric ZIP regression model to two data sets involving Alzheimer's disease and meth lab seizures. Lastly, Section 2.5 is an appendix where proofs of theorems and additional numerical work is displayed.

2.2 Estimation of Semiparametric Regression Models

Semiparametric ZIP Regression Model

Suppose that $\{(Y_i, \boldsymbol{X}_i, \boldsymbol{W}_i)\}_{i=1}^n$ is a random sample, where the conditional distribution of $Y|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{W}_i = \boldsymbol{w}_i$ is distributed as

$$Y|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{W}_i = \boldsymbol{w}_i \sim \pi(\boldsymbol{w}_i)\mathrm{I}\{y_i = 0\} + (1 - \pi(\boldsymbol{w}_i))p(y_i; \mu_i(\boldsymbol{\beta})), \tag{2.12}$$

$p(\cdot; \mu_i(\boldsymbol{\beta}))$ is the Poisson mass function with mean $\mu_i(\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, and the zero-inflation probability $\pi(\boldsymbol{w}_i)$ are nonparametric smooth functions of the covariates $\boldsymbol{w}_i$. For simiplicity, we'll assume that $\boldsymbol{w}_i \in \mathbb{R}^1$, and write w_i instead of $\boldsymbol{w}_i$. The methodology can be easily extended to multivariate $\boldsymbol{w}_i$, but one needs to be cognizant of the "curse of dimensionality" [136]. To overcome the curse of dimensionality, one typically assumes the structure of $\pi(\boldsymbol{w})$ is additive, which is referred to as a *generalized additive model* (GAM). In other words, one assumes

$$\pi(\boldsymbol{w}) = \alpha_0 + \pi(w_1) + \pi(w_2) + \cdots + \pi(w_q),$$

where $\mathbb{E}(w_i)$ to ensure identifiability. Another method to overcome the curse of dimensionality is the *partially linear model*, which supposes $\pi(\boldsymbol{w})$ can be written as a linear function of covariates plus a nonparametric smooth function $g(\cdot)$ of a continuous covariate T; i.e.,

$$\pi(\boldsymbol{w}, t) = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w} + g(t).$$

Identifiablity

Identifiability is a paramount concern in mixture models. For valid interpretation of parameters, the following notion of identifiability is needed.

Definition 2.2.1. *Suppose $Y \sim f(y; \boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$ and parameter space $\boldsymbol{\Theta}$. The model $f(y; \boldsymbol{\theta})$ is said to be* identifiable *if for any $\boldsymbol{\theta}_1$,$\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$*

$$f(y; \boldsymbol{\theta}_1) = f(y; \boldsymbol{\theta}_2) \implies \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2. \tag{2.13}$$

In other words, two different combinations of parameters can't give rise to the same likelihood. For more details on identifiability of finite mixture models, see [137]. More details on identifiability of mixtures-of-regressions models can be seen in [138].

In regards of identifiability of the ZIP regression model, [139] proved the identifiability of aforemetioned model for the univariate case of $\mu = \beta_0 + \beta_1 x$ and for any smooth function $\pi(x)$. Under more general conditions, the identifiability of the semiparametric ZIP regression model can be proved with the aid of a theorem from [140], who studied identifiablity of nonparametric and semiparametric mixtures of GLMs. The ZIP regression model can be viewed as a mixture of GLMs, where we view the degnerate component as a Poisson distribution with rate parameter $\mu \equiv 0$.

For model 2.12, the conditions for identifiablity can be seen in Theorem 2.1. The proof is given in the Appendix.

Theorem 2.2.1. *Model 2.12 is identifiable if all of the following conditions are satisfied:*

1. *The domain $\mathcal{X}$ of $\boldsymbol{x}$ contains an open set in $\mathbb{R}^p$, and the domain $\mathcal{W}$ of w has no isolated points.*

2. *$\pi(w) > 0$ are continuous functions.*

3. *The parametric ZIP model $\pi I\{y = 0\} + (1 - \pi)p(y; \mu)$ is identifiable.*

Conditions (1) and (2) are assumptions on the domain space and parameter space, thus, all that is left to show is condition (3).

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \pi(\cdot))$, and let $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ denote the likelihood function for the sample $\{(Y_i, \boldsymbol{X}_i, W_i)\}_{i=1}^n$. Since $\pi(\cdot)$ is nonparametric, ℓ is not ready for optimization. To learn $\pi(\cdot)$, local-likelihood methodology will be employed. First, local likelihood regression will be reviewed.

Local Likelihood

Assume $Y_i|X_i = x_i \sim f(y; \theta(x_i))$ for $i = 1, \ldots, n$, where $f(\cdot; \theta(x_i))$ is a density (or pmf) in the exponential family with canonical parameter $\theta(x)$. We assume $\theta(x)$ is d-times differentiable fucnction of x. Often, it is assumed that $d = 1$ or $d = 2$. Then, by Taylor's Theorem, $\theta(x_0)$ can be well approximated by a d-degree polynomial for points in the domain "close" to x_0. More formally, if for $x \in \mathbb{R}$ such that $|x - x_0| < h$, where h is sufficiently small, it follows that

$$\begin{aligned} \theta(x) &\approx a_0 + a_1(x - x_0) + \frac{1}{2}a_2(x - x_0)^2 + \ldots \frac{1}{d}(x - x_0)^d \\ &= \boldsymbol{a}^{\mathrm{T}}\boldsymbol{A}(x - x_0), \end{aligned} \tag{2.14}$$

where $\boldsymbol{a} = (a_0, \ldots, a_d)^{\mathrm{T}}$ and $\boldsymbol{A}(v) = (1, v, \ldots, \frac{1}{d}v^d)^{\mathrm{T}}$ is the polynomial basis. Then, at a fixed point x_0 in the predictor space, define the smoothed-likelihood at x_0 as

$$\ell_{x_0}^S(\boldsymbol{a}) = \sum_{i=1}^n w_i \log f\big(y_i|\theta(x) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{A}(x - x_0)\big), \tag{2.15}$$

where the $w_i = h^{-1}K\big(\frac{x_i - x_0}{h}\big)$ are weights and $K(\cdot)$ is some kernel function with bandwidth h. Let $\widehat{\boldsymbol{a}}$ be the maximizer of 2.15. Then, the local likelihood estimate of

Table 2.1: Common Kernel Functions

Kernel	$K(u)$
Uniform	$\frac{1}{2}\mathrm{I}\{\lvert u\rvert \leq 1\}$
Epanechnikov	$\frac{3}{4}(1-u^2)\mathrm{I}\{\lvert u\rvert \leq 1\}$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$

$\theta(x_0)$ is defined as

$$\widehat{\theta}(x_0) = \widehat{a}_0. \tag{2.16}$$

In essence, the local likelihood at x_0 is a weighted average of the log densities of the samples such that observations that are closer to x_0 have more weight or influence over the estimate of $\theta(x_0)$. The *bandwidth* h controls the size of the neighborhood around x_0. To learn the function $\theta(x)$, typically we estimate $\theta(\cdot)$ at a set of grid points $\mathcal{Z} = \{z_1, \ldots z_N\}$, and then estimate $\theta(x)$ for $x \notin \mathcal{Z}$ by linearly interpolating.

As in linear regression, a higher degree polynomial leads to estimates with less bias, but higher variability in the estimate [141]. Therefore, lower-order first or second degree approximations are preferred. Local constant approximations, also referred to the Nadaraya-Watson estimate, as exhibit low variance and are compuationally simple, but can suffer from boundary bias or more generally, bias in regions where the data is sparse. The issue can be mitigated by choosing a proper bandwidth. The local linear or local quadratic approximates correct the boundary/sparsity issues, but with more computational burden. In our work, we will use the local constant approximation, and investigate proper bandwidth selection.

Common choices of the kernel or weight function are the Uniform, Epanechnikov, and Gaussian; see Table 2.1. The choice of kernel function in kernel regression is typically inconsequential, but the choice of the *bandwidth* (or *smoothing*) parameter h is critical [142]. If h is too small, the bias will be small, but the estimated regression function will have high variability, leading to spurious features in the curve. On the other hand, if h is too large, the estimated function will exhibit low variance, but the bias will be high, which could lead to missing interesting features. Typically, bandwidths are chosen by minimizing some criterion. For example, one common measure is the *integrated square error* (ISE), which is

$$\mathrm{ISE}(h) = \int_{\mathbb{R}} \left(\widehat{m}_h(x) - m(x)\right) f_X(x) dx, \tag{2.17}$$

where $\widehat{m}_h(\cdot)$ is the estimate of the true regression function $m(x)$, and $X \sim f_X(x)$. So,

ISE is the average mean-square error, where the expectation is taken with respect to X. However, ISE is a random variable since it is a function of Y, and thus is difficult to minimize. Therefore, we then consider the *mean integrated squared error* (MISE), which is defined as

$$\begin{aligned} \text{MISE}(h) &= \mathbb{E}\left(\text{ISE}(h)\right) \\ &= \int \cdots \int \text{ISE}(h) f(x_1, \ldots, x_n, y_1, \ldots, y_n) dx_1 \ldots dx_n dy_1 \ldots dy_n \end{aligned} \tag{2.18}$$

where the expectation is taken with respect to the joint distribution of the observed sample $\{(X_i, Y_i)\}_{i=1}^n$. MISE is not a random variable, and a minimizer of 2.18 can be derived; see [128] and [142]. When taking the Taylor series expansion of 2.18, the quanity contains higher-order terms that are on the order of $o(h^4)$ and $o((nh)^{-1})$. The typical asymptotic conditons of kernel regression are $h \to 0$ and $nh \to \infty$, and thus these higer order terms vanish, and so we can then consider the *Asymptotic Mean Integrated Square-Error* (AMISE). The formula for the optimal bandwidth chosen by AMISE can be seen in [143]. But, the minimizer of AMISE depends on unkown quantities, such as $m''(x)$, which depends on the function we are estimating. This then leads to the *plug-in approach* for bandwidth selection, where these unknown quantities are replaced by estimates, such as those obtained by a polynomial regression fit.

Another common method for choosing the bandwidth, which is what we will employ, is cross-validation. Let $L(\widehat{m}_h(\cdot), m(\cdot))$ be a loss function (ex. L^2) for quantifying the loss between the true function m and its estimate $\widehat{m}$. Ideally, we would like to find $\underset{h}{\text{argmin}}\, L(\widehat{m}_h(\cdot), m(\cdot))$. But, m is unknown, and thus m is replaced with the responses Y_i's, which is then can be seen of a measure of how well $m_h(x)$ predicts Y. The issue is that many times this quantity can be made arbitrarily small by choosing $\widehat{m}_h(\cdot)$ to interpolate each of the Y_i's. For example, if we consider the *average square error loss* (ASE), $L(\widehat{m}_h, \boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^n \{m_h(X_i) - Y_i\}^2$, we can then choose m_h to interpolate each Y_i's since the $\widehat{m}_h$ is trained on the Y_i's. To remedy this dilemma, we will employ K-fold *cross-validation* (CV).

Cross-validation is performed as follows:

1. Consider a grid of candidate bandwidths $\mathcal{H} = \{h_1, \ldots h_M\}$ where $h_1 < h_2 < \cdots < h_m$.

2. Partition the whole data set $\mathcal{D}$ randomly into a training set $\mathcal{R}_j$ and test set $\mathcal{T}_j$ for $j = 1, \ldots K$ such that $\mathcal{T}_j \cap \mathcal{T}_{j'} = \emptyset$ for $j \neq j'$.

3. For $h \in \mathcal{H}$, do:

a) For $j = 1, \ldots, K$, do:

i. Train the model on $\mathcal{R}_j$, and obtain $\widehat{m}_h(x)$.

ii. Output:

$$CV^{(j)}(h) = \sum_{l \in \mathcal{T}_j} L(\widehat{m}_h(X_l), Y_l).$$

b) Output :

$$CV(h) = \sum_{j=1}^{K} CV^{(j)}(h).$$

4. Output :

$$\widehat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, CV(h).$$

Typically K is set to 5 or 10. Cross-validation can exhibit high variability, so it is recommended to repeat the procedure 30 to 50 times, and then taking the average of the CV bandwidths. It can be shown that the $\widehat{h} = O(n^{-1/5})$, which doesn't meet the under-smoothing requirements for the asymptotic theory. A suggested adjusted under-smoothed bandwidth by [144] is $\widetilde{h} = \widehat{h} \times n^{-2/15} = O(n^{-1/3})$. In this chapter, under-smoothing, CV-smoothing, and over-smoothing will be investigated.

Bandwidth selection is among the most highly debated issues in semiparametric regression. Plug-in approaches are computationally simple and stable, but requires estimating unknown quantities about the function of estimation. CV approaches optimizes the bandwidth on independent validation sets, but is computationally expensive and can exhibit high variability. For a more in depth discussion on the difficulty of choosing bandwidths, refer to [145].

Another approach to bandwith selection can be found in the computer vision subfield known as scale space theory. From the perspective of [146] and [147], adjusted h is "like adjusting the focus on a camera" [148]. According to [148], "A larger h gives a macroscopic view of the surface, showing only large-scale features, while a small h gives a zoomed-in view to show small scale features". In summary, the scale space view of bandwidth selection looks at a range of bandwidths to help determine which features are consistantly present across multiple bandwidths. [148] developed a local likelihood regression "SIgnificant ZERo crossings of derivatives" (SiZer) map applying some of the scale space ideas to local generalized linear models. Another recent development in the vein of scale space theory is taking a confidence interval approach to choosing h; see [149].

Estimation

The challenge in estimating 2.12 is that it that it contains both parametric and non-parametric functions. Similar to the estimation in [122], we propose a one-step *backfitting* algorithm for estimation of $\boldsymbol{\beta}$ and $\pi(\cdot)$. Backfitting estimation alternates between steps of local estimation and global estimation. The algorithm was first proposed by [150], and then further studied by [151]. Before we propose a backfitting algorithm for estimation of 2.12, let us review the backfitting algorithm in the context of the *partially-linear model* (PLM), which is similar to our model in that it contains both parametric and nonparametric parts. The summary below follows that from Chapter 7.1 of [142]

The PLM can be written as

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + m(T) + \epsilon,$$

where ϵ is an error term with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) < \infty$. Moreover, $(\boldsymbol{X}^{\mathrm{T}}, T)^{\mathrm{T}} \perp \epsilon$. Then,

$$Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = m(T) + \epsilon,$$

which implies

$$\mathbb{E}\left(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\middle|\, T\right) = m(T). \tag{2.19}$$

Suppose we had an inital estimate $\widehat{\boldsymbol{\beta}}$, say from a linear regression of Y on $\boldsymbol{X}$. Plugging in $\widehat{\boldsymbol{\beta}}$, we can now estimate $m(T)$ by running a nonparametric regression of $Y - \boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ on T; call this $\widehat{m}(T)$. Then, fixing the nonparametric estimate $\widehat{m}(T)$, consider

$$\mathbb{E}(Y - \widehat{m}(T)|\boldsymbol{X}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}. \tag{2.20}$$

Thus, we can update $\widehat{\boldsymbol{\beta}}$ by running a linear regression of $Y - \widehat{m}(T)$ on $\boldsymbol{X}$. We then alternate between nonparametric estimation of $m(\cdot)$ and parametric estimation of $\boldsymbol{\beta}$ until convergence.

Returning to the backfitting estimation of 2.12, let $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ denote the likelihood of the sample, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \pi(\cdot))$. That is,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log\left[\pi(w_i)\mathrm{I}\{y_i = 0\} + (1 - \pi(w_i))p(y_i; \mu_i(\boldsymbol{\beta}))\right]. \tag{2.21}$$

2.21 is not ready for optimization as it contains the nonparametric function $\pi(\cdot)$. Let $\mathcal{Z} = \{z_1, \ldots z_N\}$ be a set of grid points for local estimation. Define the smoothed

likelihood at $z \in \mathcal{Z}$ as

$$\ell_z^{S_1}(\pi_0, \boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} K_h(w_i - z) \log \Big[\pi_0 \mathrm{I}\{y_i = 0\} + (1 - \pi_0) p(y_i; \mu_i(\boldsymbol{\beta})) \Big], \tag{2.22}$$

where $K_h(t) = h^{-1}K(t/h)$ is a rescaling of the kernel function with bandwidth h. Here π_0 is the local constant in local likelihood regression to be estimated; i.e., $\widehat{\pi}(z_0) = \widehat{\pi}_0$, where $\widehat{\pi}_0$ is the maximizer of 2.22 . Maximization of 2.22 is difficult, so we'll use an "EM-like" algorithm similar to that seen in Section 1.3. The term "EM-like" is used because we are no longer maximizing a true likelihood function, but instead a weighted version of 2.21. Similar to Section 1.3, we can calculate a complete-data local likelihood that is analagous to (1.11), but the kernel weights are incorperated as an additional weighted component for the regressions. The complete-data local likelihod at $z \in \mathcal{Z}$ is

$$\begin{aligned} \ell_{C,z}^{S_1}(\pi_0, \boldsymbol{\beta}; \boldsymbol{y}) = &\sum_{i=1}^{n} K_h(w_i - z) \big(r_i \log(\pi_0) + (1 - r_i) \log(1 - \pi_0) \big) \\ &+ \sum_{i=1}^{n} K_h(w_i - z)(1 - r_i) \big(y_i \log(\mu_i(\boldsymbol{\beta})) - \mu_i(\boldsymbol{\beta}) \big). \end{aligned} \tag{2.23}$$

Thus, seperation of the Poisson component and binary regression component in the complete-data local likelihood is achieved, and we can optimize π_0 and $\boldsymbol{\beta}$ seperately. Note here that the estimate of $\boldsymbol{\beta}$ is a function of z. The EM-like algorithm for maximizing 2.22 is as follows:

Step 1: For $t = 0, 1, 2 \ldots$, til convergence, do:

1. For all observations $i = 1, \ldots n$, update the posterior membership probabilities

$$r_i^{(t+1)} = \begin{cases} \frac{\pi^{(t+1)}(w_i)}{\pi^{(t+1)}(w_i) + (1 - \pi^{(t+1)}(w_i)) \exp\{-\mu_i(\boldsymbol{\beta}^{(t+1)}(w_i))\}} & y_i = 0 \\ 0 & y_i > 0. \end{cases} \tag{2.24}$$

2. For all $z \in \mathcal{Z}$ do:

   a) Update $\pi^{(t+1)}(z)$ by

$$\pi^{(t+1)}(z) = \frac{\sum_{i=1}^{n} r_i^{(t+1)} K_h(w_i - z)}{\sum_{i=1}^{n} K_h(w_i - z)}, \tag{2.25}$$

   and $\boldsymbol{\beta}^{(t+1)}$ via weighted Poisson regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ with weight matrix $Q^{(t+1)} = \mathrm{diag}\{(1 - r_1^{(t+1)}) K_h(w_1 - z), \ldots, (1 - r_n^{(t+1)}) K_h(w_n - z)\}$.

3. For all observations $i = 1, \ldots, n$, update $\pi^{(t+1)}(w_i)$ and $\boldsymbol{\beta}^{(t+1)}(w_i)$ by linearly interpolating $\pi^{(t+1)}(z)$ and $\boldsymbol{\beta}^{(t+1)}(z)$ for $z \in \mathcal{Z}$.

4. Output $\pi^{(t+1)}(w_i)$ and $\boldsymbol{\beta}^{(t+1)}(w_i)$ for all $i = 1, \ldots, n$.

Call these intial estimates $\widetilde{\boldsymbol{\beta}}(w)$ and $\widetilde{\pi}(w)$. Asymptotic properties of these estimates will be discussed later in this section. Since $\boldsymbol{\beta}$ is a global parameter, $\widetilde{\boldsymbol{\beta}}(w)$ will not have $\sqrt{n}$-consistancy. Thus, we now want to perform a global estimation of $\boldsymbol{\beta}$.
Step 2: Fix $\widetilde{\pi}(z)$ at its current estimate. Define then global likelihood as

$$\ell_2(\boldsymbol{\beta}; \boldsymbol{y}, \widetilde{\pi}(w)) = \sum_{i=1}^{n} \big[\widetilde{\pi}(w_i)\mathrm{I}\{y_i = 0\} + (1 - \widetilde{\pi}(w_i))p(y_i; \mu_i(\boldsymbol{\beta}))\big]. \tag{2.26}$$

We can then run an EM algorithm to maximize 2.26.
For $t = 0, 1, \ldots,$ til convergence, do:

1. Update the posterior memberships:

$$r_i^{(t+1)} = \frac{\widetilde{\pi}(w_i)}{\widetilde{\pi}(w_i) + (1 - \widetilde{\pi}(w_i)) \exp(-\mu_i(\boldsymbol{\beta}^{(t+1)}))}. \tag{2.27}$$

2. Update $\boldsymbol{\beta}^{(t+1)}$ via weighted Poisson regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ with weight matrix $\boldsymbol{Q}^{(t+1)} = \mathrm{diag}\{1 - r_1, \ldots, 1 - r_n\}$. Output $\boldsymbol{\beta}^{(t+1)}$

Call this global estimate $\widehat{\boldsymbol{\beta}}$. This is the final estimate of $\boldsymbol{\beta}$, and we will discuss asymptotic properties later in this section.

Now, we end with a final estimate of the zero-inflation probabilities. Fix $\widehat{\boldsymbol{\beta}}$ at the estimate from Step 2. We now seek to maximize

$$\ell_{C,z}^{S_3}(\pi_0; \boldsymbol{y}, \widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} K_h(w_i - u) \log\big[\pi_0 + (1 - \pi_0)p(y_i; -\mu_i(\widehat{\boldsymbol{\beta}}))\big]. \tag{2.28}$$

Maximization of 2.28 is completed via the following "EM-like" algorithm analagous to Step 1:
Step 3: For $t = 0, 1 \ldots,$ til convergence, do:

1. For all observations $i = 1, \ldots, n$, update the posterior probabilities

$$r_i^{(t+1)} = \frac{\pi^{(t+1)}(w_i)}{\pi^{(t+1)}(w_i) + (1 - \pi^{(t+1)}(w_i)) \exp(-\mu_i(\widehat{\boldsymbol{\beta}}))}. \tag{2.29}$$

2. For $z \in \mathcal{Z}$, do:

a) Update $\pi^{(t+1)}(z)$:

$$\pi^{(t+1)}(z) = \frac{\sum_{i=1}^{n} K_h(w_i - z) r_i^{(t+1)}}{\sum_{i=1}^{n} K_h(w_i - z)}. \tag{2.30}$$

3. Update $\pi^{(t+1)}(w_i)$ for all observations $i = 1, \ldots, n$ by interpolating $\pi^{(t+1)}(z)$ for $z \in \mathcal{Z}$.

4. Output $\pi^{(t+1)}(w_i)$ for $i = 1, \ldots, n$.

Call these final estimates $\widehat{\pi}(z)$. Asymptotic properties will be discussed later in this section.

Theoretical Properties of Estimation

Now, we examine some theoretical properties of the proposed estimation from the "EM-like" algorithm, as well as *ascent-type* properties from using the local likelihood functions as surrogates. The proofs follow *mutatis mutandis* from [122] and [128], and are included in the Appendix. We first discuss asymptotic properties of the estimators at each step of the backfitting process.

Asymptotic Properties of Estimators

The regularity conditions for the proofs are given in the Appendix. They are not necessarily the weakest conditions possible for sufficiency, but help to facilitate the proofs. Before we discuss asymptotic normality of estimators, we lay out some notation. Let $\{(W_i, \boldsymbol{X}_i, Y_i)\}$ be a random sample from the population $(W, \boldsymbol{X}, Y)$. Define,

$$q_\theta(\boldsymbol{\theta}; w, \boldsymbol{x}, y) = \frac{\partial \ell(\boldsymbol{\theta}; y)}{\partial \boldsymbol{\theta}},$$
$$q_{\theta\theta}(\boldsymbol{\theta}; w, \boldsymbol{x}, y) = \frac{\partial^2 \ell(\boldsymbol{\theta}; y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}}.$$

Analagously, we can define q_π, $q_{\pi\pi}$, q_β, $q_{\beta\beta}$, and $q_{\beta\pi}$. Moreover, define

$$G(w) = \mathbb{E}\big[q_\theta(\boldsymbol{\theta}(w), \boldsymbol{X}, Y) | W = w\big],$$
$$\Gamma(w) = \mathbb{E}\big[q_\pi(\boldsymbol{\theta}(w); \boldsymbol{X}, Y) | W = w\big].$$

Furthermore, define the localized versions of the Fisher information matrices as

$$\begin{aligned}
\mathcal{I}_\theta(w) &= -\mathbb{E}\big[q_{\theta\theta}(\boldsymbol{\theta}(w); W, \boldsymbol{X}, Y)|W = w\big], \\
\mathcal{I}_\beta(w) &= -\mathbb{E}\big[q_{\beta\beta}(\boldsymbol{\theta}(w); W, \boldsymbol{X}, Y)|W = w\big], \\
\mathcal{I}_\pi(w) &= -\mathbb{E}\big[q_{\pi\pi}(\boldsymbol{\theta}(w); W, \boldsymbol{X}, Y)|W = w\big], \\
\mathcal{I}_{\beta\pi}(w) &= -\mathbb{E}\big[q_{\beta\pi}(\boldsymbol{\theta}(w); W, \boldsymbol{X}, Y)|W = w\big],
\end{aligned}$$

and,

$$\omega(w, \boldsymbol{x}, y) = \mathcal{I}_{\beta\pi}(w)\psi(w, \boldsymbol{x}, y),$$

where $\psi(w, \boldsymbol{x}, y)$ is the first element of $\mathcal{I}_\theta(w)q_\theta(\boldsymbol{\theta}; w, \boldsymbol{x}, y)$. Let $W \sim g(w)$, with support $\mathcal{W}$. Then, for the first step of the EM-like algorithm, we have the following convergence property.

Theorem 2.2.2. *Fix $z_0 \in \mathcal{W}$, and let $\widetilde{\boldsymbol{\theta}}(z_0) = (\widetilde{\pi}(z_0), \widetilde{\boldsymbol{\beta}}(z_0))$ be the maximizer of 2.22 at the fixed point z_0. Let $\boldsymbol{\theta}(z_0)$ be the true value of $\boldsymbol{\theta}$ at z_0. Assume $n \to \infty$, $h \to 0$, and $nh \to \infty$. Then, under the regularity conditions in the Appendix,*

$$\sqrt{nh}\{\widetilde{\boldsymbol{\theta}}(z_0) - \boldsymbol{\theta}(z_0) - \boldsymbol{b}(z_0)h^2 + o(h^2)\} \xrightarrow{L} \mathcal{N}(\mathbf{0}, g^{-1}(z_0)\mathcal{I}_\theta^{-1}(z_0)v), \tag{2.31}$$

where $v = \int K^2(t)dt$,

$$b(z_0) = \mathcal{I}_\theta^{-1}(z_0)\left[\frac{G'(z_0)g'(z_0)}{g(z_0)} + \frac{1}{2}G''(z_0)\right]\mu_2, \tag{2.32}$$

and $\mu_2 = \int t^2K(t)dt$ is the second moment of the kernel function. Moreover, $G'(\cdot)$ and $G''(\cdot)$ refer to the first and second derivatives of each component of the vector $G(\cdot)$ with respect to w.

As alluded to before, estimating $\boldsymbol{\beta}$ locally will lead to a loss in efficiency. Therefore, we investigate the limiting distribution of the maximizer of 2.26.

Theorem 2.2.3. *Fix $\widetilde{\pi}(w)$ at the initial estimate from Step 1 of the backfitting procedure. Let $\boldsymbol{\beta}$ be the true population value, and let $\widehat{\boldsymbol{\beta}}$ be the maximizer of 2.26. Assume $nh^4 \to 0$ and $nh^2\log(h^{-1}) \to \infty$. Then, under the regularity conditions given in the Appendix, it follows*

$$\sqrt{n}\{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\} \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{B}^{-1}\boldsymbol{\Sigma}\boldsymbol{B}^{-1}), \tag{2.33}$$

where $\boldsymbol{B} = \mathbb{E}(\mathcal{I}_\beta(w))$, *and*

$$\boldsymbol{\Sigma} = Var\left[\frac{\partial \ell(\pi(W), \boldsymbol{\beta}; W, \boldsymbol{X}, Y)}{\partial \boldsymbol{\beta}} - \omega(W, \boldsymbol{X}, Y)\right].$$

Finally, we investigate the final estimate of the mixing proportions, $\widehat{\pi}(w)$.

Theorem 2.2.4. *Fix* $z_0 \in \mathcal{W}$, *and let* $\widehat{\pi}(z_0)$ *be the maximizer of 2.28. Let* $\pi(z_0)$ *be the true population value. Assume* $n \to \infty$, $h \to 0$, *and* $nh \to \infty$. *Then, under the regularity conditions given in the Appendix, we have*

$$\sqrt{nh}\{\widehat{\pi}(z_0) - \pi(z_0) - b^*(z_0)h^2 + o(h^2)\} \xrightarrow{L} \mathcal{N}(0, g^{-1}(z_0)\mathcal{I}_\pi^{-1}(z_0)v), \tag{2.34}$$

where again $v = \int K^2(t)dt$. *The asymptotic bias* $b^*(z_0)$ *is given by*

$$b^*(z_0) = \mathcal{I}_\pi^{-1}(z_0)\Big\{\frac{\Gamma'(z_0)g'(z_0)}{g(z_0)} + \frac{1}{2}\Gamma''(z_0)\Big\}\mu_2,$$

where $\mu_2 = \int t^2 K(t)$.

Finally, we can show that that the final backfitting estimate $\widehat{\pi}(w)$ is at least as efficient and has bias that is less than or equal to that of $\widetilde{\pi}(w)$. We state this as a theorem for reference.

Theorem 2.2.5. *Fix* $z_0 \in \mathcal{W}$. *Then the asymptotic variance and asymptotic bias of* $\widehat{\pi}(z_0)$ *are both less than or equal to the asymptotic bias and asymptotic variance of* $\widetilde{\pi}(z_0)$.

A final note about selection of h is that $h = h_n$ needs to have large enough order with respect to n for the asymptotic theory to hold. According to [142], the CV bandwidth is $h = o(n^{-1/5})$. For Theorem 2.2.3, it is assumed that $nh^4 \to 0$. Then, $n(\widehat{h}^{-1/5})^4 = O(n) \times O(n^{-4/5}) = O(n^{1/5}) \neq o(1)$. Therefore, $\widehat{h}$ does not meet the asymptotic requirements for Theorem 2.2.3. But, the undersmoothed bandwidth, $\widehat{h} \times n^{-2/15} = O(n^{-1/3}) = o(1)$, and so the undersmoothed bandwidth meets the asymptotic requirements.

Ascent Properties of EM-like Algorithm

Classical EM algorithms are known to possess the *ascent property* [118].

Definition 2.2.2. *Let $\boldsymbol{\theta}$ be a vector of parameters, and $\ell_o(\boldsymbol{\theta})$ be the observed (marginal) likelihood. The* ascent property *for EM algorithms is the property that*

$$\ell_o(\boldsymbol{\theta}^{(t+1)}) \geq \ell_o(\boldsymbol{\theta}^{(t)}) \tag{2.35}$$

for all iterations $t = 0, 1, \ldots,$ of the algorithm.

In other words, the objective function is monotone increasing in each iteration of the algorithm. For the EM-like algorithm presented, it is probably too strong to claim that $\ell(\boldsymbol{\theta})$ is monotone increasing at each iteration, although [121] noticed in their simulations that the observed likelihood was always monotone in the EM-like algorithm. But, we can replace the classical ascent property with some weaker claims about montonicity.

Theorem 2.2.6. *The following statements hold:*

1. *For the EM-like algorithm in Step 1, assume that $nh \to \infty$ as $n \to \infty$ and $h \to 0$. Suppose $t \to \infty$. Fix $z_0 \in \mathcal{W}$. Then,*

$$\liminf_{n\to\infty} \quad \mathbb{P}\left\{ n^{-1}\left[\ell_{z_0}^{S_1}(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}) - \ell_{z_0}^{S_1}(\boldsymbol{\theta}^{(t)}; \boldsymbol{y})\right] \geq 0 \right\} = 1. \tag{2.36}$$

2. *For Step 2 of the algorithm, $\ell_2(\boldsymbol{\beta}^{(t+1)}; \boldsymbol{y}) \geq \ell_2(\boldsymbol{\beta}^{(t)}; \boldsymbol{y})$.*

3. *At Step 3 of the algorithm, for any $z_0 \in \mathcal{W}$, $\ell_{z_0}^{S_3}(\pi^{(t+1)}; \boldsymbol{y}) \geq \ell_{z_0}^{S_3}(\pi^{(t)}; \boldsymbol{y})$.*

In interpretation, (1) implies that when the sample size is large, that the ascent property holds in $\ell_z^{S_1}(\boldsymbol{\theta}^{(t)})$ at a fixed $z_0 \in \mathcal{W}$ at large iterations of t. We refer to this as the *asymptotic ascent property*. Property (2) follows directly from the theory of ordinary EM algorithms since it is a parametric estimation. Again, property (2) means that the parametric likelihood $\ell_2(\boldsymbol{\beta}; \boldsymbol{y})$ has the ascent property. Finally, property (3) implies that the estimates from the EM-like algorithm are monotone increasing in t for $\ell_{C,z_0}^{S_3}(\pi^{(t)}; \boldsymbol{y})$ at any fixed $z_0 \in \mathcal{W}$.

2.3 Inference

After model fitting, we are interested in conducting inference. The first question of interest may be "Given the subject is at risk for the event (i.e., observation comes from count component), what factors lead to increase (decrease) in the number of incidents?". Moreover, interest may lie in estimating the probability of a subject not

being at risk (i.e., degenerate component) for a given w, or we may want to study how the mixing proportion changes with w. Finally, one may ask "Is the semiparametric model an improvement over the parametric model?".

With respect to studying how covariates affect μ and π, construction of confidence intervals will be essential. Notice in the asymptotic results for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\pi}$ that the asymptotic variance depend on unknown and intractable quantities. Therefore, the bootstrap will be utilized for confidence interval estimation. First, we review the parametric bootstrap.

Bootstrap Intervals

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta(x)$, where $F_\theta(x)$ is known except for θ. Suppose we have an estimate $\widehat{\theta}(X_1, \ldots, X_n)$ of θ. Now we are interested in the limiting distribution of

$$C_n = \sqrt{n}(\widehat{\theta} - \theta)$$

since this usually has a familar limiting distribution, such as the normal distribution. The common issue is that even though we know the limiting distribution of C_n is say normal, the asymptotic variance is intractable. Therefore, an approximation of the distrubtion function for C_n is needed. The parametric bootstrap alogorthim is:
For $t = 1, \ldots, B$, where is sufficiently large, do:

1. Generate $X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} F_{\widehat{\theta}}$.
2. Estimate $\widehat{\theta}_t^*$ from $X_1^*, \ldots, X_n^*$.

Let $C_n^* = \sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$, and we approximate the limiting distribution of F_{C_n} by $\widehat{F}_{C_n^*}$, where

$$\widehat{F}_{C_n^*}(x) = \frac{1}{B}\sum_{t=1}^{B} \mathrm{I}\{\sqrt{n}(\widehat{\theta}_t^* - \widehat{\theta}) \leq x\}.$$

It can be shown that

$$\sup_{x \in \mathbb{R}} |\widehat{F}_{C_n^*} - F_{C_n}| \xrightarrow{p} 0.$$

Therefore, the limiting distribution function of C_n can be approximated by $\widehat{F}_{C_n^*}$. Hence, $\mathrm{Var}(\widehat{\theta})$ can be approximated by $\widehat{\mathrm{Var}}(\widehat{\theta}) = \frac{1}{B}\sum_{t=1}^{B}(\widehat{\theta}_t^* - \widehat{\theta})^2$.

After estimating the variance or limiting distribution of $\widehat{\theta}$, we now seek interval estimates for $\widehat{\theta}$. For our proposed model, we know the limiting distribution is asymtotically normal, and therefore an approximate $1-\alpha$ confidence interval (CI) for $\widehat{\theta}$ is

$$\widehat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta})}. \tag{2.37}$$

[152] investigate bootstrap intervals of the form 2.37 in the nonparametric regression setting. The authors' simulation studies show that when the asymptotic bias is not addressed, the coverage rates can be significantly smaller than nominal. The authors argue mitigating this problem by using the bootstrap to obtain an estimate bias, or by using an locally"optimal" bandwidth $h(w)$ at different grid points, which they call *local adaptive smoothing*. Another possible solution proposed by [153] is to use an "oversmoothed" bandwidth to fit the overall model, and then obtain a bias estimate of the regression via bootstrap or by calculation. Then, using the bias adjusted estimate of the curve, perform the bootstrap using the "optimal" bandwidth to obtain estimates of variability at the grid points. Moreover, [154] proposes to use a single under-smoothed bandwidth for the overall fit and the bootstrap fits to construct confidence bands.

In addition to the normal-based intervals, *percentile intervals* can also be employed. Order the bootstrap samples

$$\widehat{\theta}^*_{(1)} \leq \widehat{\theta}^*_{(2)} \leq \cdots \leq \widehat{\theta}^*_{(B)}. \tag{2.38}$$

It was discussed above that we can view the resampling estimates, $\widehat{\theta}^*_j$, as (approximate) samples of the limiting distriubtion of $\widehat{\theta}$. Therefore, an approximate $1-\alpha$ CI is given by

$$\left[\widehat{\theta}^*_{(\alpha_1)}, \widehat{\theta}^*_{1-(\alpha_2)}\right],$$

where $\alpha_1 + \alpha_2 = \alpha$, $1-\alpha_2 > \alpha_1$ for $\alpha_1, \alpha_2 > 0$. Moreover, $\widehat{\theta}^*_{(\gamma)}$ denotes the sample γ^{th} quantile.

Lastly, another bootstrap interval method is the *Bias Corrected* (BC) intervals. BC intervals correct the for the bias of $\widehat{\theta}$. A $1-\alpha$ BC interval is defined as

$$\left[\widehat{\theta}^*_{(\alpha_1)}, \widehat{\theta}^*_{(\alpha_2)}\right],$$

where

$$\alpha_1 = \Phi(2\widehat{z}_0 + z_{(\alpha)}),$$
$$\alpha_2 = \Phi(2\widehat{z}_0 + z_{(1-\alpha)}).$$

Here, $\Phi(\cdot)$ is the distribution function of the standard normal distribution, and $z_{(\alpha)}$ is the α^{th} quantile of the standard normal distribution. Moreover, the bias correction

factor, $\widehat{z}_0$, is defined as

$$\widehat{z}_0 = \Phi^{-1}\Big(B^{-1}\sum_{t=1}^{B} \mathrm{I}\{\widehat{\theta}^*_{(t)} < \widehat{\theta}\}\Big).$$

BC intervals are part of a more general class of intervals called *Bias Corrected and Acclerated* (BC_a) intervals. The a stands for the accelaration constant, which corrects for the common issue that the $\mathrm{Var}(\widehat{\theta})$ depends on the true value of θ [155]. The quantity a can be difficult to estimate, and thus we will not consider BC_a intervals in our simulations. However, the normal-based, percentile, and BC intervals will be studied.

The resampling procedure for the semiparametric ZIP regression model is similar to the parametric bootstrap. The procedure is as follows:
For $t = 1, \ldots, B$, do:

1. Generate $Y_1^*|W_1, \boldsymbol{X}_1, \ldots, Y_n^*|W_n, \boldsymbol{X}_n \sim \mathrm{ZIP}(\widehat{\pi}(w_i), \mu_i(\widehat{\boldsymbol{\beta}}))$.
2. Calculate $\widehat{\boldsymbol{\beta}}^*_t$ and $\widehat{\pi}^*_{(t)}(\cdot)$.

Output the three aformentioned confidence intervals.

Generalized Likelihood Ratio Test

In addition to constructing confidence intervals for the parameters, we may be interested in testing whether the semiparametric model provides a better fit than the fully parametric model. Formally, we want to test

$$\begin{aligned} H_0 &: \ \pi(w) \in \mathcal{M}_\theta \\ H_1 &: \ \pi(w) \notin \mathcal{M}_\theta, \end{aligned} \tag{2.39}$$

where $\mathcal{M}_\theta$ are a class of parametric models indexed by θ. A special case of this hypothesis is

$$\begin{aligned} H_0 &: \ \pi(w) = c \\ H_1 &: \ \pi(w) \neq c, \end{aligned} \tag{2.40}$$

where c is a constant with $0 < c < 1$. In other words, "Does the mixing proportion depend on w?".

[156] argue that that likelihood ratio test (LRT) can be employed to test 2.39 if the nonparametric function is replaced with a quality estimater, such as a sieve MLE

or a local regression estimate. Similar to the classic parametric LRT, the test statistic is

$$\lambda_n(h) = 2\{\ell(H_1) - \ell(H_0)\}, \tag{2.41}$$

where $\ell(H_1)$ and $\ell(H_0)$ are the likelihoods under the alternative and null hypothesis, respectively. Typically, under H_0,

$$r\lambda_n \overset{D}{\approx} \chi^2_{\mu_n}$$

for a sequence $\mu_n \to \infty$ and a constant r, such that

$$(2\mu_n)^{-1/2}(r\lambda_n - \mu_n) \xrightarrow{L} \mathcal{N}(0,1),$$

where μ_n and r are free of any nuisance parameters [157]; i.e., the null hypothesis estimates of $\boldsymbol{\beta}$ and $\pi(w)$. The authors dub this result as the *Wilks's phenomenon*. However, there is no well-defined degrees of freedom under H_1 since it is a semi-parametric model. Moreover, calculating μ_n and r can be intractable, but since the limiting distribution of $r\lambda_n$ does not depend on nuisance parameters, a parametric bootstrap can be utilized to approximate the limiting distribution of λ_n under H_0, and then a bootstrap p-value can be calculated for the hypothesis test. The bootstrap procedure is as follows:

1. For $Y_1|W_1, \boldsymbol{X}_1, \ldots, Y_n|W_n, \boldsymbol{X}_n \sim \mathrm{ZIP}(\pi(w_i), \mu_i(\boldsymbol{\beta}))$, fit both the semiparametric and fully parametric models, and then calculate λ_n. Call the MLEs of $\pi(w)$ and $\boldsymbol{\beta}$ under the null hypothesis $\overline{\pi}(w)$ and $\overline{\boldsymbol{\beta}}$, respectively.
2. For $t = 1, \ldots, B$, do:
   a) Generate $Y_1^*|W_1, \boldsymbol{X}_1, \ldots Y_n^*|W_n, \boldsymbol{X}_n \sim \mathrm{ZI}(\overline{\pi}(w_i), \mu_i(\overline{\boldsymbol{\beta}}))$.
   b) Fit both the fully parametric and semiparametric models, and compute λ_t^*.
3. Compute the bootstrap p-value, $p_B = B^{-1}\sum_{t=1}^{B} \mathrm{I}\{\lambda_t^* > \lambda_n\}$. If $p_B < \alpha$, for a suitably chosen type-1 error rate α, then reject H_0.

A common issue encountered with the bootstrap LRT is negative bootstrap LRT statistics, which is technically non-sensical since the null hypothesis is nested within the alternative. Two possible reasons for this issue is the non-negligible smoothing bias of the semiparametric estimation of $\pi(\cdot)$, and the fact that the parametric model converges at the rate $\sqrt{n}$, whereas the semiparametric model converges at the slower

rate of $\sqrt{nh}$. To mitigate the later issue, [142] recommends a bias-adjusted LRT statistic. [157] also provides a bias-corrected LRT statistic.

Another matter in the LRT is choice of smoothing parameter, h. In general, "smaller values of h will be more powerful against less smooth alternatives, and larger values of h is more powerful against smoother alternatives."[157]. [157] further discuss methods of bandwidth choice such as grid search or the ad hoc bandwidth, $\widehat{h} \times n^{-1/45}$, where $\widehat{h}$ is the optimal bandwidth with respect to some criterion. In our simulations, we study power with the under-smoothed, CV, and over-smoothed bandwidths. Further discussion by eminent statisticians on the generalized LRT can be seen in the discussion section of [157].

2.4 Simulation Studies

To examine the performance of the proposed model, two simulation studies were conducted. The second simulation study is provided in the Appendix. For π and $\boldsymbol{\beta}$, accuracy of point estimates and coverage probabilities for interval estimates were studied. Two samples sizes were examined : $n \in \{200, 400\}$. The data were generated from the ZIP regression model with a single covariate, X, where $X \sim \text{Unif}(0, 1)$. The true $\boldsymbol{\beta} = (.5, 1.5)^{\text{T}}$, and the true zero-inflation probability is

$$\pi(x) = .2 + .75 \sin(\pi x).$$

To select the optimal bandwidth, 50 independent data sets were generated from the aforementioned model. The CV process was applied to each data set, and the optimal bandwidth, $\widehat{h}_k^*$, was recorded. Then, the bandwidth chosen for the optimal bandwidth employed for the simulations is $\widehat{h} = \frac{1}{50}\sum_{k=1}^{50} \widehat{h}_k^*$. Three bandwidths were examined: $n^{-2/15}\widehat{h}$, $\widehat{h}$, and $2\widehat{h}$. These bandwidths correspond to under-smoothed, optimally smoothed, and over-smoothed, respectively.

Accuracy of $\widehat{\boldsymbol{\beta}}$ was judged by *mean-square error* (MSE); i.e., $\text{MSE}(\widehat{\beta}_k, \beta_k) = \mathbb{E}\big\{(\widehat{\beta}_k - \beta_k)^2\big\}$ for $k = 1, 2$. The accuracy of $\widehat{\pi}(x)$ was determined by the *root average square errors* (RASE) measure, which is defined as

$$\text{RASE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\widehat{\pi}(x_i) - \pi(x)]^2}.$$

For comparison, we include the fully parametric model in the simulation. Numerical results can be seen in 2.2. Graphical displays of 2.2 can be seen in **??**. With respect

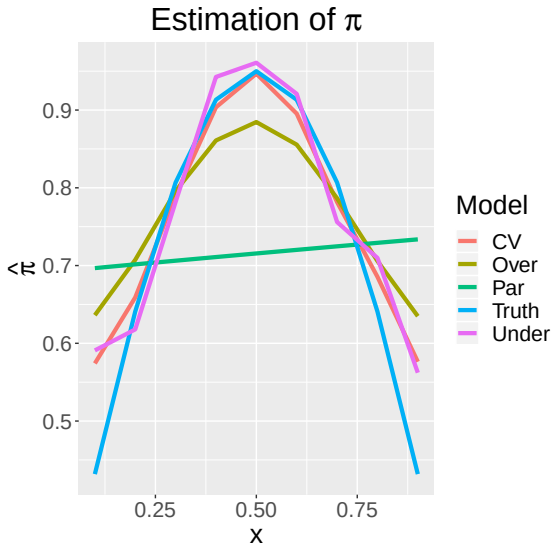


Figure 2.1: Estimated Curve for $\pi(x)$

Table 2.2: Examination of Average MSE and Average RASE.

	Bandwidth($n = 200$)				Bandwidth($n = 400$)			
MSE	.04	.08	.16	PAR	.03	.07	.15	PAR
β_0	0.240	0.0733	0.026	0.026	0.095	0.018	0.016	0.016
β_1	0.131	0.070	0.046	0.046	0.057	0.026	0.026	0.026
				RASE$_\pi$				
π	0.107	0.097	0.133	0.233	0.090	0.084	0.130	0.232

to estimation of $\boldsymbol{\beta}$, observe that that the MSE is smaller as the bandwidth increases. The oversmothed bandwidth, which is the most "similar" to the parametric model, has similar MSE for the estimation of $\boldsymbol{\beta}$. Also, for $n = 400$, the MSE for the optimal bandwidth is comparable to that of the oversmooted bandwidth and parametric model. However, estimation of π is substantially poorer for the oversmoothed bandwidth and the parametric model. Instead, the optimally smoothed bandwidth had the smallest RASE for both sample sizes. Estimation of the curve for $\pi(x)$ from one Monte-Carlo sample can be see Figure 2.1.

Before we discuss confidence interval performance, we briefly discuss bootstrap calibration. Rarely is the nominal coverage level equal to the actual coverage of a bootstrap CI. Ideally, we would like to seek a mapping $\alpha \mapsto \lambda$ such that coverage probability of the interval $(\widehat{\theta}_\lambda, \widehat{\theta}_{1-\lambda})$ is equal to the nominal rate $1 - \alpha$. One method for finding λ is the *double bootstrap* method, which is a specific method of *bootstrap*

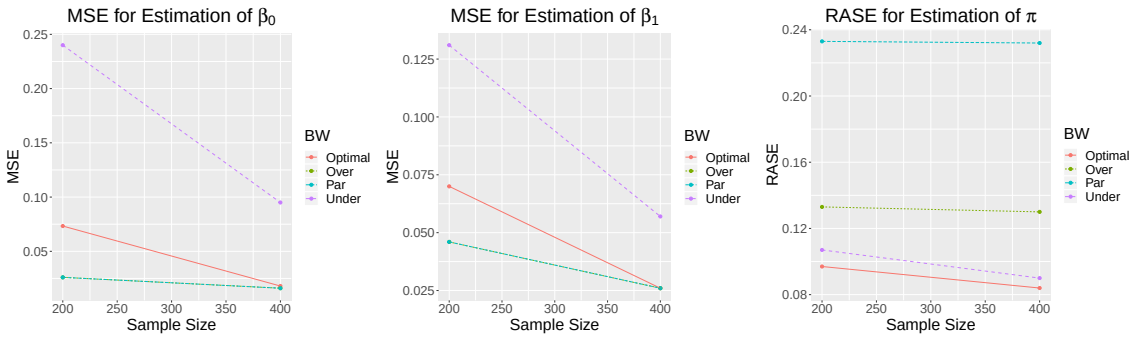
calibration. Double bootstrap methods are computationally expensive, typically requiring at least B^2 resampling procedures. For more details on the double bootstrap, see [158] Other methods include calibrating the quantiles, location, and standard error of $\widehat{\theta}$; see [159] . Bootstrap calibration is beyond the scope of this work, and instead, ad hoc methods will be employed.

The coverage results for $\boldsymbol{\beta}$ can be seen in 2.3. We report the Z-intervals with the standard error estimated with the bootstrap. We can see that for all sample sizes and bandwidths, the coverage is near the nominal level. It is suprising to see that there is a slight decrease in coverage as n increases, although, [122] reported similar results in their semiparametric mixture of regressions article. Note in Theorem 2.2.3 that $nh^4 \to 0$, and therefore, the undersmoothed bandwidth is the closest to meeting that requirement. Therefore, the undersmoothed bandwidth will provide the best coverage.

The coverage results for π can be seen in 2.4. Coverage probabilites were examined at the grid points $x = .1, .2, \ldots, .9$. Observe that the Z-intervals (top entry in each cell) are typically very conservative for the x in the interior; i.e., $.2 \leq x \leq .8$. Moreover, the intervals have poor coverage near the boundary of the predictor space (i.e., $x = .1$ or $x = .9$), regardless of the bandwidth. Thus, we conclude that the Z-intervals are not reliable.

In contrast, the percentile intervals (middle row in each cell) are a bit more reliable. But, they still suffer from poor coverage near the edge of the predictor space, and the CV bandwidth is unreliable at various points of x. Hence, while the percentile intervals are an improvement over the Z-intervals, the coverage probabilities are still unsatisfactory.

Lastly, it seems that the BC intervals are the most reliable of the three. The type-



MSE for β_0 MSE for β_1 RASE for π

Figure 2.2: Plot of Average MSE and RASE.

1 error rate was ad-hoc calibrarted at $\gamma = \alpha/2 = .025$. For the under-smoothed and CV bandwidth, the coverage rates are near nominal levels for all x. This underscores the importance of adjusting for the bias in the estimator of π. Another observation is that for $x = .5$, the coverage rates are below nominal, even for the undersmoothed and CV bandwidth. Notice that the true mixing proportion at $x = .5$ is $\pi(.5) = .95$, and therefore, we are getting close to the boundary of the parameter space. It is of interest on how to calibrate the interval using more rigorous methodology. Graphical display of the coverage probabilities for the BC intervals when $n = 400$ can be seen in 2.3.

One final observation is that, in general, the oversmoothed bandwidth provides poor coverage at most values of x, regardless of the interval methodology being employed. This highlights the importance in choosing a reasonable smoothing parameter. The oversmoothed bandwidth, while exhibiting less variance in estimation of π, has large bias, which yield poor coverage.

Finally, we examine the power of the bootstrap LRT. The power function is estimated under a sequence of local alternatives

$$\begin{aligned} H_0 &: \pi(x) = .2 \\ H_1 &: \pi(x) = .2 + .75\delta \sin(\pi x)/\sqrt{nh}, \end{aligned}$$

where $\delta \in \{0, .5, 1, \ldots, 2.5\}$. Here, δ expresses the amount of weight on the non-linear component in H_1, and $\sqrt{nh}$ is analagous to the sample size in local regression. Note that if $\delta = 0$, then the alternative collapses into the null. Moreover, three type-I error rates, $\alpha \in \{.01, .05, .1\}$, were examined. For each n, δ, and h, $M = 500$ Monte Carlo data sets were generated with the mixing proportion defined in H_1, and from each

Table 2.3: Coverage Results for $\boldsymbol{\beta}$

Parameter	95%	95%
	$n = 200$,$h = .04$	$n = 400$,$h = .04$
β_0	96.38	93.00
β_1	96.00	95.16
	$n = 200$,$h = .08$	$n = 400$,$h = .08$
β_0	94.67	91.90
β_1	95.43	93.32
	$n = 200$,$h = .15$	$n = 400$,$h = .15$
β_0	94.29	90.32
β_1	96.00	93.49

Table 2.4: Coverage Rates for Intervals for π. The three numbers in each cell represent the coverage for the Z, percentile, and BC interval, respectively, for a value of x and h.

	$n = 200$			$n = 400$		
h	.04	.08	.16	.03	.07	.15
	99.81	89.90	0.00	40.40	0.50	0.00
.1	97.52	83.24	5.33	96.99	62.94	0.00
	92.57	94.29	32.38	95.99	89.82	3.01
	100.00	100.00	100.00	100.00	100.00	90.32
.2	96.57	95.43	82.10	95.49	96.49	66.78
	92.95	92.57	91.43	94.82	95.66	90.48
	100.00	100.00	100.00	100.00	100.00	100.00
.3	97.71	93.52	75.24	96.66	73.79	55.43
	95.24	95.24	88.38	96.66	95.99	83.97
	98.48	100.00	99.81	99.83	100.00	56.43
.4	94.29	86.86	2.67	97.16	53.26	0.00
	89.71	91.62	30.67	94.82	93.49	1.17
	99.81	100.00	80.95	99.83	100.00	0.00
.5	95.05	79.05	0.00	95.99	70.12	0.00
	89.33	91.81	1.90	91.15	89.65	0.00
	100.00	100.00	100.00	100.00	100.00	4.51
.6	97.28	84.95	1.14	98.00	87.81	0.00
	93.71	92.57	24.19	95.66	93.32	1.17
	100.00	100.00	100.00	100.00	100.00	100.00
.7	98.28	92.95	61.33	96.99	96.83	31.05
	94.10	93.71	85.90	97.16	95.49	74.29
	100.00	100.00	100.00	100.00	100.00	78.46
.8	97.90	96.95	89.53	96.99	73.46	84.14
	94.86	95.05	94.67	95.99	95.99	94.66
	99.81	26.48	0.00	63.94	0.00	0.00
.9	97.90	86.29	9.52	96.83	73.46	0.50
	95.05	96.00	42.86	96.33	95.16	6.84

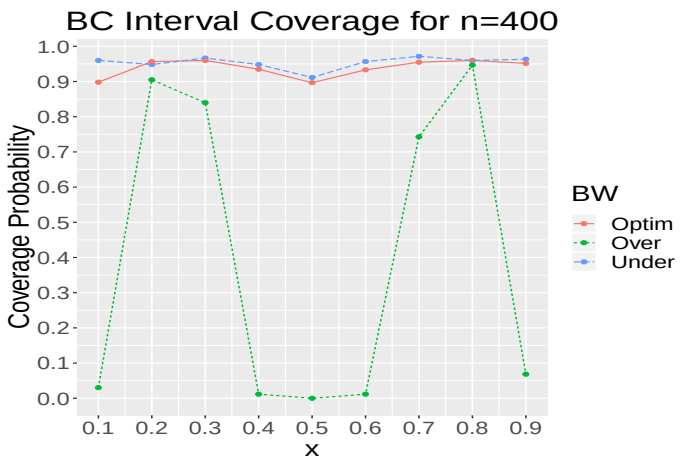


Figure 2.3: BC Intervals for π when $n = 400$

data set Monte Carlo rep, $B = 500$ bootstrap data sets were generated. Then, for each data set and its corresponding 500 bootstrap data sets, the bootstrap LRT was performed, along with the calculation bootstrap p-value. The acceptance/rejection was then recorded for each α. The power function, $p_{H_1}(\delta, \alpha, n, h)$, was then estimated by $\widehat{p}_{H_1}(\delta, \alpha, n, h) = M^{-1} \sum_{m=1}^{M} \mathrm{I}\{\text{Rejection of } H_0 \text{ for the } M^{th} \text{ Monte Carlo Rep}\}$.

A density plot of the bootstrap LRT statistics can be seen in Figure 2.4. The simulation results are presented in 2.5. Observe that for both n and all h, the simulated type-I error rates are close to the nominal level. Moreover, the undersmoothed bandwidth provides the highest power across all combination of n and α, followed by the optimal bandwidth. The oversmoothed bandwidth provides the poorest power, which coincides with the poor aforementioned estimation results. We can see all bandwidths yield consistant tests; namely, the power approaches 1 as we deviate further from the null hypothesis. Lastly, there is a spurious decrease in the power curve for the undersmoothed bandwidth as δ goes from 2.5 to 3. The reason for that is when $\delta = 3$, the mixing proportions generated from H_1 are close to the boundary of the parameter space. For example, when $h = .04$, $n = 200$,$\delta = 3$, $\pi(.5) = .9955 \approx 1$, where $\pi(x)$ is of the form in the alternative hypothesis. Therefore, the spurious decrease is most likely due to the mixing proportions approaching degeneracy.

2.5 Real Data Analysis

Alzheimer's Data

The data set are from the Biologically Resilient Adults in Neurological Studies (BRAiNS) study, and are provided by Dr. Dave Fardo and Dr. Pete Nelson from the University

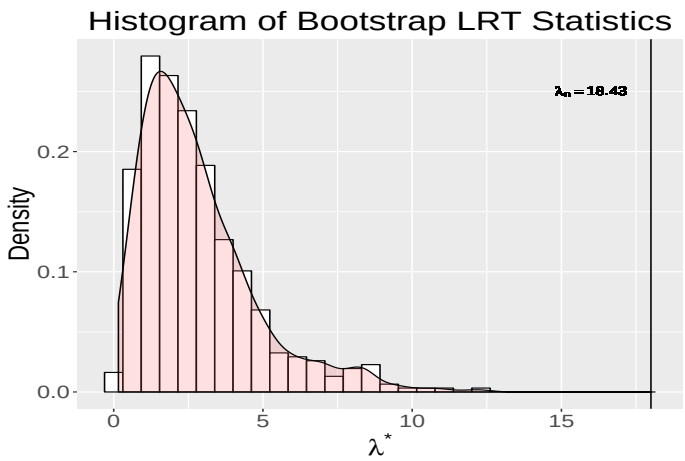
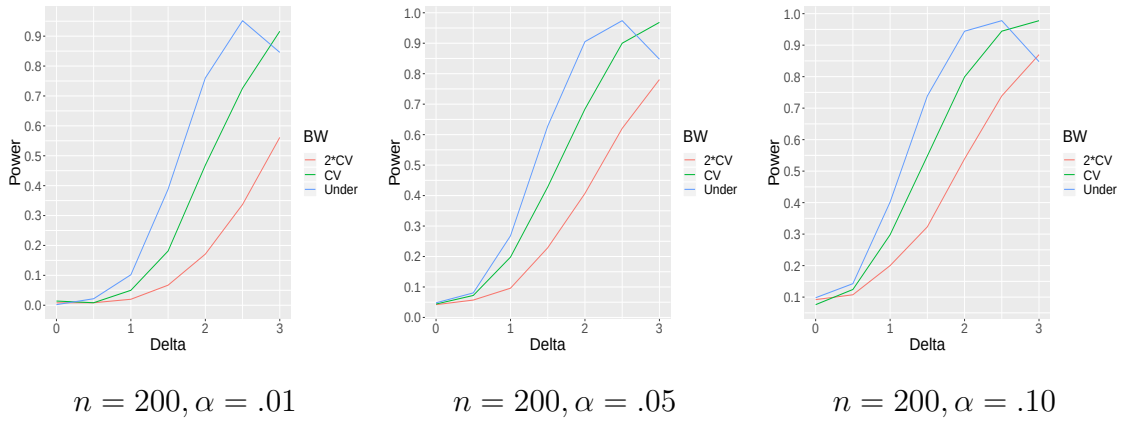


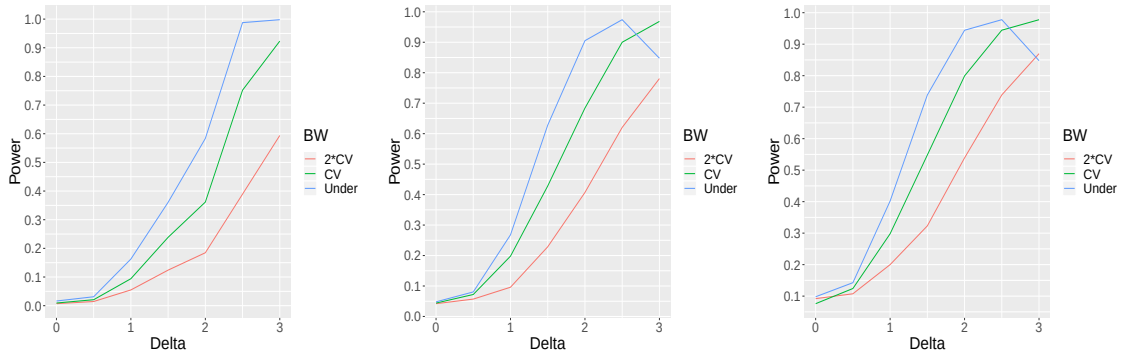
Figure 2.4: Histogram of bootstrap LRT statistics for a single Monte-Carlo replicate

of Kentucky's Sanders-Brown Center on Aging. The goal of the of the BRAiNS study is to further understand neurodegenerative diseases such as Alzheimer's by examing quantitative brain pathologies in the elderly. There are 126 participants to date who were autopsied, during which the number of TDP-43 (TAR DNA-binding protein 43) inclusions in the subiculum, Triiodothyronine level (T3), age at death, sex, and the patient's score of the Mini-Mental State Examination (MMSE) were recorded. TDP-43 is the chief protein that is present in the brain for many types of neurodegnerative diseases, such as frontotemporal dementia (FTD) and amyotrophic lateral scleros (ALS). An inclusion is an aggregates of proteins - in this case, TDP-43. The subiculum is a subregion of the Hippocampus, which plays a critical role in memory and attention control. Thus, higher counts of TDP-43 inclusions represent higher amounts of cognitive degeneration and imperative. T3 is a hormone produced by the thyroid gland, and has been shown to be positively associated with longer completion times on the Trail Making Test and Tower of London Test. Lastly, MMSE is a cognitive exam to test patients memory and critical thinking skills. The score ranges from 0 to 30, where a score of 0-12, 13 -20, 20-24, and >24 indicate severe dementia, moderate dementia, mild dementia, and no dementia, respectively. Here, we are interested in modeling the number of TDP-43 inclusions in the subiculum by the covarates T3, age at death, sex, and MMSE score.

A histogram of inclusion counts, truncated at five inclusions, can be seen in Figure 2.6. Overall, $87/126 \approx 69\%$ of the patients had zero inclusions, with a maximum count of 57 inclusions. The parametric and semiparametric model were fit with T3, age at death, sex along with the interaction of age and sex , with MMSE in the zero-inflation state. Bandwidth selection for the semiparametric model was performed



$n = 200, \alpha = .01$ $n = 200, \alpha = .05$ $n = 200, \alpha = .10$



$n = 400, \alpha = .01$ $n = 400, \alpha = .05$ $n = 400, \alpha = .10$

Figure 2.5: Simulated Power Functions of the LRT

Table 2.5: Comparison of Beta Coefficients and log-likelihoods.

Model	Intercept	T3	Age	Sex	Sex*Age	Log-like
Undersmoothed	3.3645 (0.5684)	-0.0705 (0.0168)	-0.0012 (0.0064)	-2.3600 (0.8281)	-0.0325 (0.0100)	-318.6023
Optimal	3.3644 (0.5957)	-0.0705 (0.0179)	-0.0012 (0.0068)	-2.3598 (0.8749)	-0.0325 (0.0108)	-321.5853
Oversmoothed	3.3645 (0.6300)	-0.0705 (0.0172)	-0.0012 (0.0073)	-2.3598 (0.9168)	-0.0325 (0.0112)	-325.7847
Parametric	8.0840 (2.1869)	-0.0705 (0.0148)	-0.0661 (0.0265)	-2.3598 (1.2257)	0.0324 (0.0144)	-325.0125

via CV by a grid search of 20 equally spaced values between 2 and 15. The optimal bandwidth was $h_{opt} = 6.5$, but the undersmoothing ($n^{-2/15} \times h_{opt}$) and oversmoothing ($1.5 \times h_{opt}$) were also examined.

A summary of the Poisson regression coefficients, along with log-likelihoods, of the four models can be seen in 2.3. A partial dependency plot of the estimated Poisson mean against T3 by sex can be seen in Figure **??**. We see that as the level of T3 increases, the average amount of inclusions decreases. In Figure **??**, a partial dependency plot of the estimated Poisson mean against age by sex is presented. For the male group, increased age does not increase the average risk of inclusions, but there is a lower average counts of inclusions as age increases for females.

A plot of the estimated zero-inflation for the four models can be seen in Figure

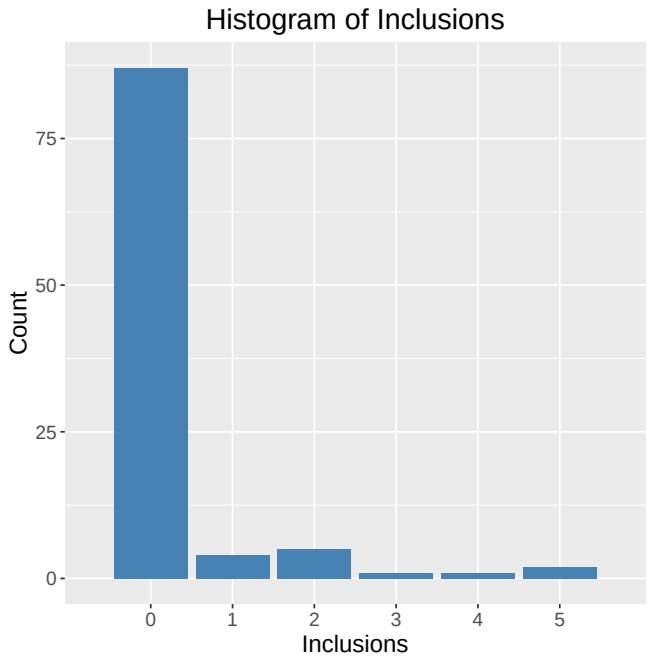


Figure 2.6: Histogram of Subiculum Inclusions. Histogram is truncated at five.

2.8. In general, the four models indicate as the exam score increases, the probability of a zero inclusion count increases. We would expect that less demented patients have smaller inclusion counts, and so the models are consistant with expectations. Note though the undersmoothed and oversmoothed bandwidths present noisy estimates of π, which could be due to having no observed MMSE scores between 8 and 10. It is well known that the Nadayara-Watson estimate performs poorly when there is "gaps" in the data. Another clinical explanation is that MMSE scores become hard to assign for severly demented patients due to their lack of response to questions. However, the oversmoothed bandwidth and parametric model is consistant with what is expected scientifically. The shape of oversmoothed and parametric curves are similar, with the main differences coming from vertical shifts. This underscores the importance of examining multiple bandwidths; namely, a bandwidth that is too small will show spurious features in the curve. This could be a case where a locally adapted bandwidth may be useful.

Kentucky Meth Lab

The data set consists of the number of clandestine lab seizures in each county of Kentucky, Louisiana, and Illinois for the years 2011, 2012, and 2013. For this analysis, we will focus on Kentucky in the year 2011. The number of seizures was obtained from

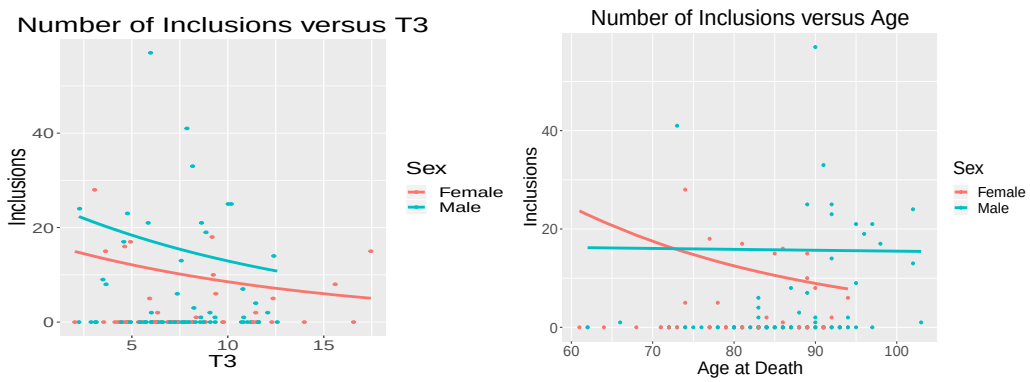


Figure 2.7: Partial dependency plot for the Inclusion data set. The left figure shows inclusions versus T3, whereas the right figure shows inclusions versus age at death.

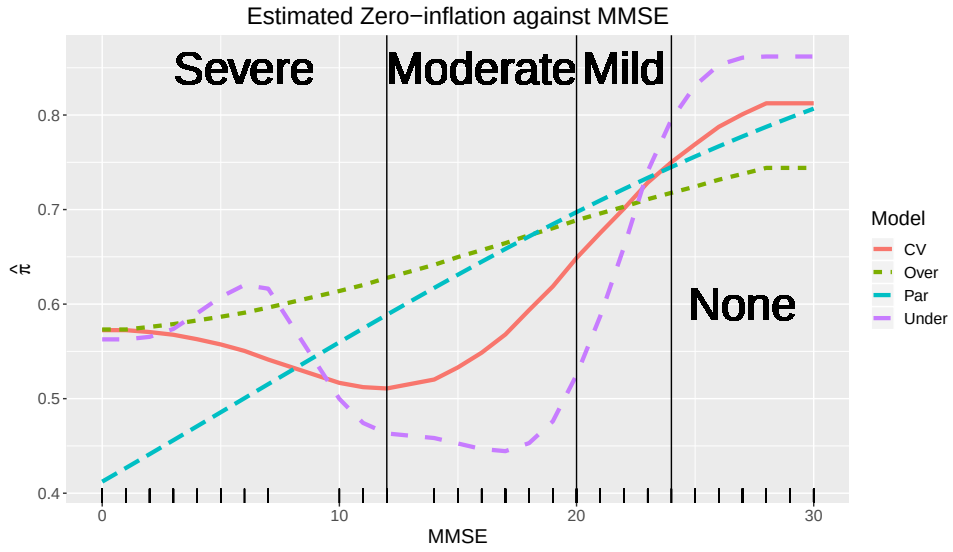


Figure 2.8: Zero-inflation Probability against MMSE

the US Drug Enforcement Administration (DEA) Clandestine Laboratory Seizure report. Interest lies in predicting the number of meth lab seizures in a county of Kentucky by socioeconomic variables, such as the median age, median income, percent poverty, etc., along with the amount of pseudophedrine (PSE) sold (in grams) per 100 people (PSE). PSE is commonly used as a sinus decongestant, but is also a main ingredient in making methamphetamine. The National Precursor Log Exchange (NPLEx) tracks the sale of all non-prescription PSE medications, which is required for all pharamacies. A histogram of the seizures can be seen in 2.9. Overall, 20% of counties had no lab seizures, with a median count of 3. The counts range from 0 to 121, with the maximum count coming from Jefferson County. Moreover, a heat map of lab counts for Kentucky can be seen in 2.10. Observe that higher counts tend to

Table 2.6: Summary of Poisson regression coefficients for Kentucky meth lab data

Model	h	β_0	β_1	β_2	β_3	ℓ_o
ZIP	*	-8.0121 (0.4450)	0.09040 (0.0097)	-0.1517 (.0078)	0.0048 (0.0009)	-1042.4940
ZINB	*	-9.2810 (2.7363)	0.1012 (0.0573)	-0.1103 (0.0382)	0.0034 (0.0037)	-371.6723
Under	28.89	-8.0016 (0.4759)	0.0949 (0.0102)	-0.1534 (0.0096)	0.0047 (0.0009)	-1022.5950
CV	54.7	-8.0030 (0.5181)	0.0948 (0.0111)	-0.1533 (0.0092)	0.0045 (0.0092)	-1043.8010
Over	75	-8.0033 (0.5092)	0.0948 (0.0110)	-0.1533 (0.0097)	0.0047 (0.0008)	-1044.3400

cluster together, which suggests a spatial model could be useful.

Lab seizure counts were predicted by median income (scaled by 1000), median age, and PSE, where the three covariates were utilized in the count component, while PSE was employed in the zero-inflation component. The three semiparametric regression models with undersmoothed, optimal, and oversmoothed bandwidths were fit. Moreover, the parametric ZIP regression model and ZINB regression model were fit. The log of the county population was applied as an offset.

A summary of the Poisson regression coefficients can be seen in 2.6. In interpretation, counties with larger average age and higher PSE sales tend to have higher counts of meth lab seizures. In contrast, wealthier counties tend to have fewer meth lab operations. The partial dependency plot of average counts against median earnings for the Poisson count components can be seen in Figure 2.11. Lastly , note that the ZINB provides a better fit to the data, which suggests the prescence of overdispersion relative to the ZIP regression model.

A plot of the zero-inflation probability against PSE sales can be seen Figure 2.12. All four models provide a similar relationship between π and PSE sold; namely, higher sales of PSE leads to lower estimates of π. In interpretation, larger sales of PSE yield low probabilites of not being at risk for the prescence of meth activity. Moreover, a (random) zero observed from the count component would mean the county is at risk for meth labs, but during the year of 2011, it was the case that we did not observe any activity. Furthermore, a zero from the degenerate component suggests that the county is at little risk for illegal meth activity.

Finally, it is of note that the PSE sales and the socio-economic variables are likely to exhibit measurement error. Therefore, it is of interest to develop a model that incorperates this source of error in the predictors.

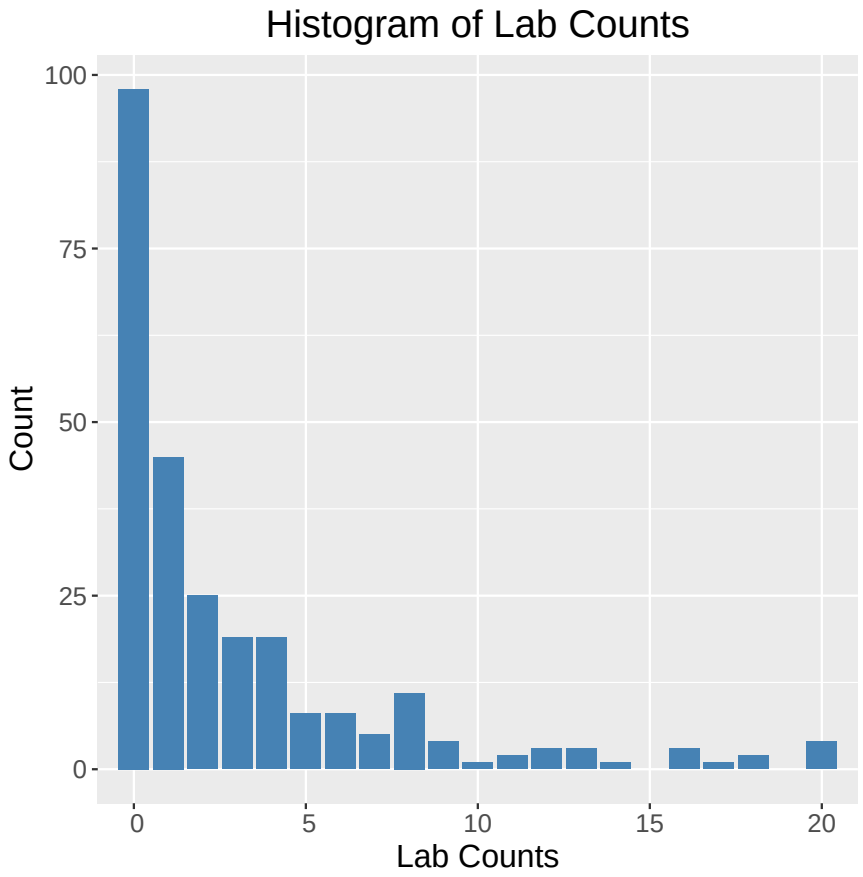


Figure 2.9: Histogram of lab seizures truncated at 20.

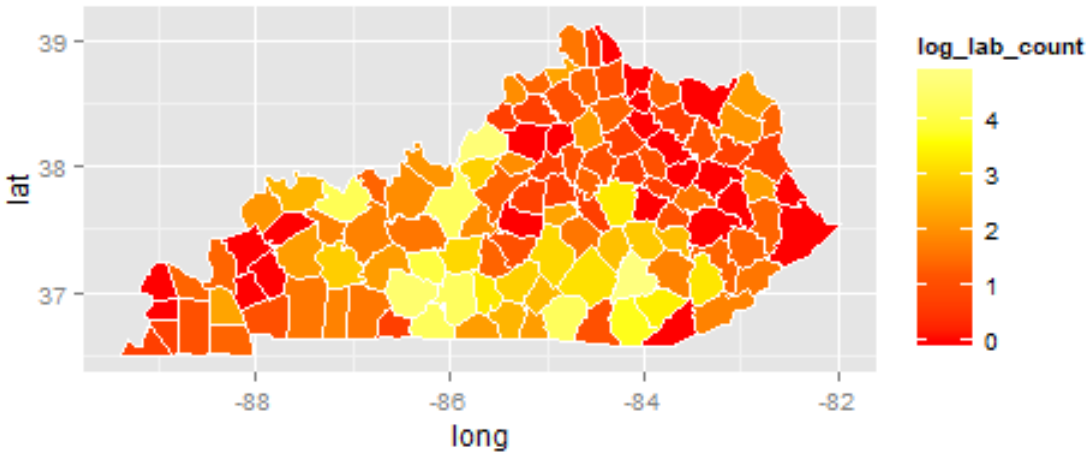


Figure 2.10: Heat map for the state of Kentucky. Yellow represents higher counts, with red representing lower counts.

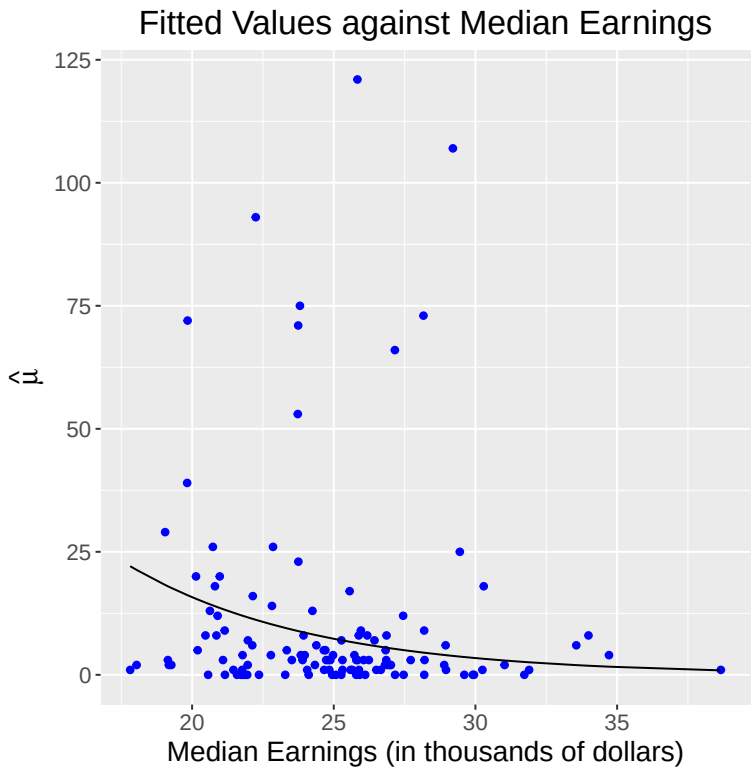


Figure 2.11: Partial dependence of lab counts on median earnings

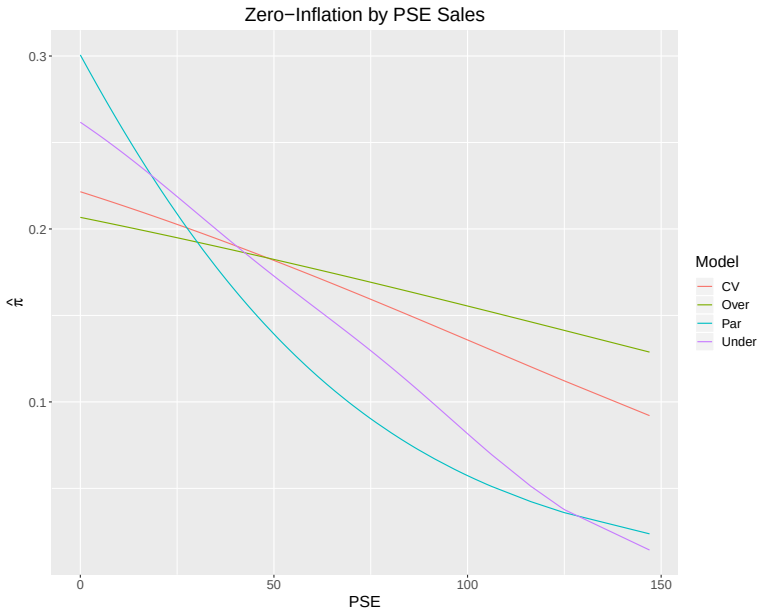


Figure 2.12: Estimated Zero-Inflation by PSE Sales

2.6 Appendix

Proof of Theorems

Proof of 2.2.1. Conditions 1 and 2 are assumptions on the predictor space and functional properties of the mixing proportions. Thus, only condition (3) needs to be shown. Rewrite the ZIP pmf as $f(y;\mu,\pi) = (1-\pi)\mathrm{I}\{y=0\} + \pi p(y;\mu)$. Suppose that $f(y;\mu,\pi) = f(y;\mu^*,\pi^*)$. Then,

$$\frac{\pi}{\pi^*} = \frac{\mathrm{I}\{y=0\} - \frac{e^{-\mu^*}(\mu^*)^y}{y!}\mathrm{I}\{y\in\mathbb{N}\}}{\mathrm{I}\{y=0\} - \frac{e^{-\mu}\mu^y}{y!}\mathrm{I}\{y\in\mathbb{N}\}} \tag{2.42}$$

Thus, setting $y=0$, 2.6 implies $\frac{\pi}{\pi^*} = \frac{1-e^{-\mu^*}}{1-e^{-\mu}}$. Moreover, setting $y=1$, we obtain $\frac{\pi}{\pi^*} = \frac{e^{-\mu^*}\mu^*}{e^{-\mu}\mu}$. Therefore,

$$\frac{e^{-\mu^*}\mu^*}{e^{-\mu}\mu} = \frac{1-e^{-\mu^*}}{1-e^{-\mu}}$$

which then implies

$$\frac{\mu^*}{\mu} = \frac{1+e^{\mu^*}}{1+e^{\mu}}$$

Therefore,

$$\frac{\left(\frac{1+\mu^*}{\mu^*}\right)}{\left(\frac{1+\mu}{\mu}\right)} = 1 \tag{2.43}$$

The function $g(\mu) = \frac{1+e^{\mu}}{\mu}$ is monotone when $\mu > 0$ since $g'(\mu) > 0$. Thus, $\frac{g(\mu^*)}{g(\mu)} = 1$, which implies $g(\mu^*) = g(\mu)$, and hence, $\mu^* = \mu$. It follows immediately that $\pi^* = \pi$. Therefore, the ZIP model is identifiable, which completes the proof. □

Asymptotic Properties of Estimators

We will assume the following regularity conditions:

1. $\pi(w)$,$\boldsymbol{\beta}(w) \in \mathcal{C}^2$.
2. $g(w) \in \mathcal{C}^2$ and $f(w) > 0$ for all $w \in \mathcal{W}$.
3. $K(\cdot)$ is symmetric about 0 and has compact support in $\mathbb{R}$.
4. The bandwidth $h \to 0$ such that as $nh \to \infty$.

Proof of 2.2.2. For convience of notation, denote $\widetilde{\boldsymbol{\theta}}(z_0) = \widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}(z_0) = \boldsymbol{\theta}$.
Note that $\widetilde{\boldsymbol{\theta}}$ satisfies

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}}\big[\ell_{z_0}^{S_1}(\boldsymbol{\theta})\big]|_{\widetilde{\boldsymbol{\theta}}} &= n^{-1}\sum_{i=1}^{n} K_h(W_i - z_0)\big\{q_\theta(\boldsymbol{\theta}; W_i, \boldsymbol{X}_i, Y_i) + q_{\theta\theta}(\boldsymbol{\theta}; W_i, \boldsymbol{X}_i, Y_i)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\big\} \\ &\quad + n^{-1}\mathcal{O}_p(\|\widetilde{\boldsymbol{\theta}}(z_0) - \boldsymbol{\theta}(z_0)\|^2) \\ &= 0.\end{aligned}$$

Define

$$\begin{aligned}W_n &= n^{-1}\sum_{i=1}^{n} K_h(W_i - z_0) q_\theta(\boldsymbol{\theta}; W_i, \boldsymbol{X}_i, Y_i) \\ \Delta_n &= -n^{-1}\sum\sum_{i=1}^{n} K_h(W_i - z_0) q_{\theta\theta}(\boldsymbol{\theta}; W_i, \boldsymbol{X}_i, Y_i).\end{aligned}$$

Then, noting that $G(z_0) = 0$,

$$\begin{aligned}\mathbb{E}(W_n) &= \mathbb{E}\big[K_h(W - z_0) q_\theta(\boldsymbol{\theta}; W, \boldsymbol{X}, Y)\big] \\ &= \mathbb{E}\bigg[\mathbb{E}\big[K_h(W - z_0) q_\theta(\boldsymbol{\theta}; W, \boldsymbol{X}, Y)|W = w\big]\bigg] \\ &= \mathbb{E}\big[K_h(W - z_0) G(z_0)\big] \\ &= \int K_h(w - z_0)\bigg[G(z_0)g(z_0) + (Gg)'(z_0)(w - z_0) \\ &\quad + \frac{1}{2}(Gg)''(z_0)(w - z_0)^2 + \dots\bigg] dw \\ &= \frac{1}{2}(Gg)''(z_0)\mu_2 h^2 + o(h^2) \\ &= \frac{1}{2}(Gg)''(z_0)\mu_2 h^2 + o(1).\end{aligned}$$

Moreover,

$$\begin{aligned}\mathrm{Var}(W_n) &= n^{-1}\bigg[\mathbb{E}\big[K^2(W - z_0) q_\theta(\boldsymbol{\theta}; W, \boldsymbol{X}, Y) q_\theta^{\mathrm{T}}(\boldsymbol{\theta}; W, \boldsymbol{X}, Y)\big] - \mathbb{E}(W_n)(\mathbb{E}(W_n))^{\mathrm{T}}\bigg] \\ &= n^{-1}\mathbb{E}\bigg[K^2(W - z)\mathbb{E}\big[q_\theta(\boldsymbol{\theta}; W, \boldsymbol{X}, Y) q_\theta^{\mathrm{T}}(\boldsymbol{\theta}; W, \boldsymbol{X}, Y)|W = w\big]\bigg] + o(1) \\ &= n^{-1}\bigg[\mathbb{E}\big[K_h^2(W - z_0)\mathcal{I}_\theta(W)\big] + o(1)\bigg] \\ &= n^{-1}\int K_h^2(w - z_0)\mathcal{I}_\theta(w) g(w) dw + o(1) \\ &= n^{-1}\int K_h^2(w - z_0)\big[\mathcal{I}(z_0)g(z_0) + o(1)\big] dw + o(1) \\ &= (nh)^{-1} g(z_0)\mathcal{I}_\theta(z_0) v + o(1).\end{aligned}$$

Furthermore, using similar arguments and Taylor Series expansions,

$$\begin{aligned}
&\mathbb{E}(\Delta_n) = \mathbb{E}\big[K_h(W - z_0)\mathcal{I}_\theta(W)\big] = \mathcal{I}(z_0)g(z_0) + o(1)\\
&\mathrm{Var}(\Delta_n(k,l)) \leq n^{-1}\mathbb{E}\left[K_h^2(W - z_0)\left\{\frac{\partial^2 \ell(\boldsymbol{\theta}; W\boldsymbol{X}, Y)}{\partial\theta_l \partial\theta_k}\right\}^2\right]\\
&= \mathcal{O}((nh)^{-1})\\
&= o(1).
\end{aligned}$$

Therefore,

$$\Delta_n = \mathcal{I}_\theta(z_0)g(z_0) + o_p(1).$$

Notice that $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = o_p(W_n)$. Then, from 2.6, it follows

$$\sqrt{nh}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\Delta_n\sqrt{nh}W_n + o_p(1) \xrightarrow{D} g^{-1}(z_0)\mathcal{I}_\theta^{-1}(z_0)\sqrt{nh}W_n + o_p(1). \tag{2.44}$$

Now, we need to show $W^* = \sqrt{nh}W_n$ is asymptotically normal. It suffices to show for any $\boldsymbol{c} \in \mathbb{R}^{p+1}$, that

$$\{\boldsymbol{c}^{\mathrm{T}}\mathrm{Var}(W_n^*)\boldsymbol{c}\}^{-\frac{1}{2}}\{\boldsymbol{c}^{\mathrm{T}}W_n^* - \boldsymbol{c}^{\mathrm{T}}\mathbb{E}(W_n^*)\} \xrightarrow{L} \mathcal{N}(0,1).$$

Let

$$\xi_i = \sqrt{h/n}K_h(W_i - z_0)\boldsymbol{c}^{\mathrm{T}}q_\theta(\boldsymbol{\theta}; W_i, \boldsymbol{X}_i, Y_i).$$

Observe that $\boldsymbol{c}^{\mathrm{T}}W_n^* = \sum_{i=1}^n \xi_i$. We now show that *Lyapunov's condition* holds. Recall the Lypunov's CLT, which is stronger than the Lindeberg-Levy CLT :

Lemma 2.6.1. *Suppose $\{X_i\}_{i=1}^n$ is a sequence of independent random variables, each with $\mathbb{E}(X_i) = \mu_i < \infty$ and $Var(X_i) = \sigma_i^2 < \infty$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. Then, if for some $\delta > 0$, the* Lyapunov's condition

$$\lim_{n\to\infty} \frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}\big[|X_i - \mu_i|^{2+\delta}\big] = 0$$

is satisfied, then

$$\frac{\sum_{i=1}^n (X_i - \mu_i)}{s_n} \xrightarrow{L} \mathcal{N}(0,1).$$

So, note that

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{c}^{\mathrm{T}}W_n^*) &= \boldsymbol{c}^{\mathrm{T}}\mathrm{Var}(W_n^*)\boldsymbol{c}\\
&= g(z_0)v(\boldsymbol{c}^{\mathrm{T}}\mathcal{I}_\theta(z_0)\boldsymbol{c}) + o(1),
\end{aligned}$$

which does not depend on n. Thus, setting $\delta = 1$, we need to show that $\sum_{i=1}^{n} \mathbb{E}(|\xi_i|^3) = n\mathbb{E}(|\xi_1|^3) = o(1)$. Since, $\mathbb{E}\big(\boldsymbol{c}^{\mathrm{T}} q_\theta(\boldsymbol{\theta}; W_i, \boldsymbol{X}_i, Y_i)\big) = M \leq \infty$, and $K(\cdot)$ has compact support,

$$\begin{aligned} n\mathbb{E}(|\xi_1|^3) &\leq nM\mathbb{E}(|K_h(W - z_0)\sqrt{h/n}|^3) \\ &= n \times (h/n)^{3/2} \times h^{-2} M^* \\ &= \mathcal{O}((nh)^{-1/2}) \\ &= o(1) \end{aligned}$$

Thus, W_n^* is asymptotically normal, and therefore

$$\sqrt{nh}\{W_n - \frac{1}{2}(Gg)''(z_0)\mu_2 h^2 + o(h^2)\} \xrightarrow{L} \mathcal{N}(\mathbf{0}, g(z_0)\mathcal{I}_\theta(z_0)v)$$

Then, applying Slutsky's Theorem, with $\Delta_n \xrightarrow{P} \mathcal{I}_\theta(z_0)g(z_0)$, it follows

$$\sqrt{nh}\{\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta} - b(z_0)h^2 + o(h^2)\} \xrightarrow{L} \mathcal{N}(\mathbf{0}, g^{-1}(z_0)\mathcal{I}_\theta^{-1}(z_0)v), \tag{2.45}$$

where

$$b(z) = \mathcal{I}_\theta^{-1}(z)\Big[\frac{G''(z)g'(z)}{g(z)} + \frac{1}{2}G''(z)\Big]\mu_2.$$

□

Proof of 2.2.3. Note that $\widehat{\boldsymbol{\beta}}$ satisfies

$$\begin{aligned} \frac{\partial \ell(\widetilde{\pi}(\cdot), \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \Bigg[n^{-1} \sum_{i=1}^{n} \frac{\partial \ell(\widetilde{\pi}(W_i), \boldsymbol{\beta}; W_i, \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\beta}}\Bigg] \\ &\quad + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \Bigg[n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \ell(\widetilde{\pi}(W_i), \boldsymbol{\beta}; W_i, \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\Bigg](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad + n^{-1} O_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) \\ &= \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} A_n + \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} B_n \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1) \\ &= 0, \end{aligned}$$

where $A_n = n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial \ell(\widetilde{\pi}(W_i); \boldsymbol{\beta}; W_i, \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\beta}}$ and $B_n = -n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \ell(\widetilde{\pi}(W_i), \boldsymbol{\beta}; W_i, \boldsymbol{X}_i, Y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}$.
Let $\widehat{\boldsymbol{\beta}}^* = \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Then,

$$0 = \widehat{\boldsymbol{\beta}}^* \big(A_n + B_n \widehat{\boldsymbol{\beta}}^*\big),$$

which implies

$$\widehat{\boldsymbol{\beta}}^* = B_n^{-1} A_n + o_p(1).$$

By the Weak Law of Large Numbers, $B_n \xrightarrow{P} -\mathbb{E}\left[\frac{\partial^2 \ell(\widetilde{\pi}(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right] = \mathbb{E}(\mathcal{I}_\beta(W_i))$.
Now taking a Taylor Series expansion of A_n around $\widetilde{\pi}(W_i)$, it follows

$$\begin{aligned}
A_n &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial \ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}} \\
&+ n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial^2 \ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}\partial\pi}\{\widetilde{\pi}(W_i)-\pi(W_i)\} + O_p(d_{1n}) \\
&= n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial \ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}} + T_{n1} + o_p(1),
\end{aligned}$$

where $d_{1n} = n^{-1/2}\|\widetilde{\pi}-\pi\|_\infty^2 = o_p(1)$. To proceed, we need the following Lemma.

Lemma 2.6.2. *Assume the regularity conditions (1)-(4) hold, and assume $nh \to \infty$ as $n \to \infty$ and $h \to 0$. Let $\sqrt{nh}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})$. Then, for all $w \in \mathcal{W}$, we have*

$$\sup_{w\in\mathcal{W}} |\widetilde{\boldsymbol{\theta}} - g^{-1}(w)\mathcal{I}_\theta^{-1}(w)\Delta_n| = O_p(h^2 + (nh)^{-1}\log^{\frac{1}{2}}(h^{-1})),$$

where $\Delta_n = \sqrt{\frac{h}{n}}\sum_{i=1}^n K_h(W_i-w)q_\theta(\boldsymbol{\theta};W_i,\boldsymbol{X}_i,Y_i)$.

For the proof of this Lemma, see [122].
Therefore,

$$\begin{aligned}
\widetilde{\boldsymbol{\theta}}(W_i) - \boldsymbol{\theta}(W_i) &= (nh)^{-1/2}\widetilde{\boldsymbol{\theta}}^* \\
&= (nh)^{-1/2}\Bigg[g^{-1}(w)\mathcal{I}_\theta^{-1}(w)\sqrt{\frac{h}{n}}\sum_{j=1}^{n} K_h(W_j-W_i)\frac{\partial\ell(\boldsymbol{\theta}(W_i);W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\theta}} \\
&+ O_p(h^2+(nh)^{-1}\log^{\frac{1}{2}}(h^{-1}))\Bigg] \\
&= n^{-1}g^{-1}(w)\mathcal{I}_\theta^{-1}(w)\sum_{j=1}^{n}\frac{\partial\ell(\boldsymbol{\theta}(W_i);W_i,\boldsymbol{X}_j,Y_j)}{\partial\boldsymbol{\theta}}K_h(W_j-W_i) \\
&+ O_p(d_{2n}),
\end{aligned}$$

where $d_{2n} = (nh)^{-1/2}h^2 + (nh)^{-1}\log^{1/2}(h^{-1})$. Note, this then implies that

$$\widetilde{\pi}(W_i) - \pi(W_i) = n^{-1}g^{-1}(W_i)\sum_{j=1}^{n}\psi(W_i,\boldsymbol{X}_j,Y_j)K_h(W_j-W_i).$$

By assumption, $nh^2(\log(h^{-1}))^{-1} \to \infty$, so

$$\begin{aligned}
n^{1/2}d_{2n} &= n^{1/2}((nh)^{-1/2}h^2 + (nh)^{-1}\log^{1/2}(h^{-1})) \\
&= h^{3/2} + n^{-1/2}h^{-1}\sqrt{\log(h^{-1})} \\
&= o_p(1).
\end{aligned}$$

Then,

$$
\begin{aligned}
T_{n1} &= n^{-1/2}\sum_{i=1}^{n}\frac{\partial^2\ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}\partial\pi}\{\widetilde{\pi}(W_i)-\pi(W_i)\}\\
&+ n^{-1/2}\sum_{i=1}^{n}\frac{\partial^2\ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}\partial\pi}\Big\{n^{-1}f^{-1}(W_i)\sum_{j=1}^{n}\mathcal{I}_\theta(W_i)\frac{\partial\ell(\pi(W_i),\boldsymbol{\beta};Z_i,\boldsymbol{X}_j,Y_j)}{\partial\boldsymbol{\beta}}K_h(W_j-W_i)\\
&+ O_p(d_{n2})\Big\}\\
&= n^{-3/2}\sum_{i=1}^{n}\frac{\partial^2\ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}\partial\pi}f^{-1}(W_i)\Big\{\sum_{j=1}^{n}\psi(W_i,\boldsymbol{X}_j,Y_j)\times K_h(W_j-W_i)\\
&+ O_p(d_{n2})\Big\}\\
&= n^{-3/2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\pi(Z_i),\boldsymbol{\beta};Z_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}\partial\pi}f^{-1}(Z_i)\psi(Z_i,\boldsymbol{X}_j,Y_j)K_h(Z_i-Z_j)+O_p(n^{1/2}h^2)\\
&= T_{n2}+O_p(n^{1/2}h^2).
\end{aligned}
$$

Let

$$
\begin{aligned}
\omega(W_j,\boldsymbol{X}_j,Y_j) &= -\mathbb{E}_{(W,\boldsymbol{X},Y)}\Big\{\frac{\partial^2\ell(\pi(W),\boldsymbol{\beta};W,\boldsymbol{X},Y)}{\partial\boldsymbol{\beta}\partial\pi}f^{-1}(W)\phi(W,\boldsymbol{X}_j,Y_j)K_h(W-W_j)\Big\}\\
&= \mathcal{I}_{\beta\pi}(W_j)\phi(W_j,\boldsymbol{X}_j,Y_j)
\end{aligned}
$$

and $T_{n3}=-n^{-1/2}\sum_{j=1}^{n}\omega(W_j,\boldsymbol{X}_j,Y_j)$. Then,

$$
\begin{aligned}
T_{n2}-T_{n3} &= n^{-3/2}\sum_{j=1}^{n}\sum_{i=1}^{n}\Big[\frac{\partial^2\ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_i,Y_i)}{\partial\boldsymbol{\beta}\partial\pi}f^{-1}(W_i)\psi(W_i,\boldsymbol{X}_j,Y_j)K_h(W_i-W_j)\\
&\quad-\omega(W_j,\boldsymbol{X}_j,Y_j)\Big]\\
&= n^{-3/2}\big[\sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{C}_{ji}\big].
\end{aligned}
$$

Note that $\mathbb{E}(\boldsymbol{C}_{ji})=0$, and that the components of $\boldsymbol{C}_{ij}$ are bounded random variables, and hence, the diagonal components of the second moment matrix $\mathbb{E}\big[(T_{n2}-T_{n3})(T_{n2}-T_{n3})^{\mathrm{T}}\big]$ are on the order of $O(Mn^{-9/4}n^2)=o(1)$ for some constant $M>0$. Therefore, Chebychev's inequality gives us that each component of $T_{n2}-T_{n3}\xrightarrow{P}0$, and so that $T_{n2}-T_{n3}\to\mathbf{0}$. Moreover, we assumed that $nh^4\to 0$, it follows that $O_p(n^{1/2}h^2)=$

$o_p(1)$. Thus,

$$\begin{aligned} A_n &= n^{-1/2}\sum_{i=1}^{n}\frac{\partial \ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_j,Y_j)}{\partial \pi} + T_{n1} + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n}\frac{\partial \ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_j,Y_j)}{\partial \pi} + (T_{n2}-T_{n3}) + T_{n3} + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n}\Big\{\frac{\partial \ell(\pi(W_i),\boldsymbol{\beta};W_i,\boldsymbol{X}_j,Y_j)}{\partial \pi} - \omega(W_i,\boldsymbol{X}_j,Y_j)\Big\} + o_p(1). \end{aligned}$$

Then, $\text{Var}(A_n) = \boldsymbol{\Sigma}$ by definition, and so

$$\mathbb{E}(A_n) = \sqrt{n}\mathbb{E}\bigg\{\frac{\partial \ell(\pi(W);\boldsymbol{\beta};W,\boldsymbol{X},Y)}{\partial \boldsymbol{\beta}} - \omega(W,\boldsymbol{X},Y)\bigg\}.$$

It can shown that the score function $\mathbb{E}\big(\frac{\partial \ell(\pi(W),\boldsymbol{\beta};W,\boldsymbol{X},Y)}{\partial \boldsymbol{\beta}}\big) = \mathbf{0}$; see [160]. Moreover,

$$\mathbb{E}\{\omega(W,\boldsymbol{X},Y)\} = -\mathbb{E}\{\mathcal{I}_{\beta\pi}(W)\psi(W,\boldsymbol{X},Y)\}.$$

Given $\mathbb{E}(\frac{\partial \ell(\pi(Z),\boldsymbol{\beta};Z,\boldsymbol{X},Y)}{\partial \boldsymbol{\beta}}) = \mathbf{0}$, it follows that

$$\begin{aligned} &\mathbb{E}\{\omega(W,\boldsymbol{X},Y)\} - \mathbb{E}_W\mathbb{E}_{\boldsymbol{X},Y|W}\{\mathcal{I}_{\beta\pi}(W)\psi(W,\boldsymbol{X},Y)|W\} \\ &\qquad = \mathbb{E}_W\mathcal{I}_{\beta\pi}(W)\mathbb{E}_{\boldsymbol{X},Y|W}\bigg[\Big(\mathcal{I}_\theta(W)\frac{\partial \ell(\pi(W),\boldsymbol{\beta};W,\boldsymbol{X},Y)}{\partial \boldsymbol{\theta}}\Big)_1|W\bigg], \end{aligned}$$

where $()_1$ denotes the first element of the vector. Note that $\mathcal{I}_\theta(W)$ is constant with respect to the inner expectation, and so if we can show that $\mathbb{E}\big(\frac{\partial \ell(\pi(W),\boldsymbol{\beta};W,\boldsymbol{X},Y)}{\partial \boldsymbol{\theta}}\big|W\big) = \mathbf{0}$, then $\mathbb{E}(\frac{\partial \ell(\pi(W),\boldsymbol{\beta};W,\boldsymbol{X},Y)}{\partial \boldsymbol{\beta}}) = \mathbf{0}$. But,

$$\begin{aligned} \mathbb{E}\Big(\frac{\partial \ell(\pi(W),\boldsymbol{\beta};W,\boldsymbol{X},Y)}{\partial \theta_k}|W\Big) &= \int \frac{\partial}{\partial \theta_k}\log dF(W,\boldsymbol{X},Y)dF_{\boldsymbol{X},Y|W} \\ &= \int \frac{\frac{\partial}{\theta_k}dF(W,\boldsymbol{X},Y)}{dF(W,\boldsymbol{X},Y)}dF_{\boldsymbol{X},Y|W} \\ &= \frac{\partial}{\theta_k}\int dF_{\boldsymbol{X},Y|W} \\ &= 0. \end{aligned}$$

Thus, $\mathbb{E}(\psi(W,\boldsymbol{X},Y)) = 0$, so that $\mathbb{E}\{\omega(W,\boldsymbol{X},Y)\} = \mathbf{0}$. Hence, $\mathbb{E}(A_n) = \mathbf{0}$, and it follows that $A_n = o_p(\sqrt{n})$.

Therefore, by the Central Limit Theorem and Slutsky's Theorem,

$$\begin{aligned}\boldsymbol{\beta}^* &= B_n^{-1}A_n + o_p(1)\\ &= B_n^{-1}\sqrt{n}(n^{-1/2}A_n) + o_p(1)\\ &\xrightarrow{L} \mathcal{N}(\mathbf{0}, B^{-1}\boldsymbol{\Sigma}B^{-1}).\end{aligned}$$

□

Proof of 2.2.4. Similar to the proof of 2.2.2, $\widehat{\pi}(z_0)$ satisfies

$$\sqrt{nh}\{\widehat{\pi}(z_0) - \pi(z_0)\} = g^{-1}(z_0)\mathcal{I}_\pi^{-1}(z_0)\widetilde{W}_n + o_p(1),$$

where

$$\widetilde{W}_n = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\pi(z_0), \widehat{\boldsymbol{\beta}}; W_i, \boldsymbol{X}_i, Y_i)}{\partial\pi}K_h(W_i - z_0).$$

Using a similar Taylor Series expansion around $\widehat{\boldsymbol{\beta}}$ in the proof of 2.2.3,

$$\begin{aligned}\widetilde{W}_n &= \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\pi(z_0), \boldsymbol{\beta}; W_i, \boldsymbol{X}_i, Y_i)}{\partial\pi}K_h(W_i - z_0)\\ &+\sqrt{\frac{h}{n}}\Bigg[\sum_{i=1}^{n}\frac{\partial^2\ell(\pi(z_0), \boldsymbol{\beta}; W_i; \boldsymbol{X}_i, Y_i)}{\partial\pi\partial\boldsymbol{\beta}}K_h(W_i - z_0)\Bigg](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1)\\ &= \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\pi(z_0), \boldsymbol{\beta}; W_i, \boldsymbol{X}_i, Y_i)}{\partial\pi}K_h(W_i - z_0) + C_n + o_p(1).\end{aligned}$$

We showed in the proof of 2.2.2 that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \mathcal{N}(\mathbf{0}, \boldsymbol{B}^{-1}\boldsymbol{\Sigma}\boldsymbol{B}^{-1})$. Moreover, $n^{-1/2}C_n \xrightarrow{P} \mathcal{I}_{\beta\pi}(z_0)g(z_0)$. Therefore,

$$\begin{aligned}C_n &= \sqrt{h}(n^{-1/2}C_n)\big(\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big)\\ &\xrightarrow{L} 0 \times \mathcal{I}_{\beta\pi}^{\mathrm{T}}(z_0)g(z_0) \times \mathcal{N}(\mathbf{0}, \boldsymbol{B}^{-1}\boldsymbol{\Sigma}\boldsymbol{B}^{-1})\\ &= 0\end{aligned}$$

Therefore, since C_n converges in distribution to a constant, $C_n = o_p(1)$.
Hence,

$$\sqrt{nh}\{\widehat{\pi}(z_0) - \pi(z_0)\} = f^{-1}(z_0)\mathcal{I}_\pi^{-1}(z_0)W_n + o_p(1),$$

where

$$W_n = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\pi(z_0), \boldsymbol{\beta}; W_i, \boldsymbol{X}_i, Y_i)}{\partial\pi}K_h(W_i - z_0) + o_p(1).$$

Using a similar calculation as in the previous proofs,

$$\mathbb{E}(W_n) = \frac{\sqrt{nh}}{2}\{\Lambda''(z_0)f(z_0) + 2\Lambda'(z_0)g'(z_0)\}h^2\mu_2 + o_p(1)$$
$$\mathrm{Var}(W_n) = \mathcal{I}_\pi(z_0)g(z_0)v_0 + o_p(1).$$

Similar to the proof of 2.2.2, we can show Lyapunov's condtion holds; thus
$W_n \xrightarrow{L} \mathcal{N}(0, \mathcal{I}_\pi(z_0)g(z_0)v_0)$.
Therefore, Slutsky's Theorem gives us

$$\sqrt{nh}\{\widehat{\pi}(z_0) - \pi(z_0) - \widehat{b}(z_0)h^2 + o(h^2)\} \xrightarrow{L} \mathcal{N}(0, g^{-1}(z_0)\mathcal{I}_\pi^{-1}(z_0)v_0)$$

where

$$\widehat{b}(z) = \mathcal{I}_\pi^{-1}(z)\mu_2\Big[\frac{g'(z)\Lambda'(z)}{g(z)} + \frac{1}{2}\Lambda''(z)\Big]$$

□

Proof of 2.2.5. Note that $\mathcal{I}_\pi(z_0)$ is the (1,1) element of $\mathcal{I}_\theta(z_0)$, and $\Lambda(z)$ is the first entry of $G(z)$. Let $\big[\mathcal{I}_\theta^{-1}(z_0)\big]_{kl}$ denote the k, l element of $\mathcal{I}_\theta^{-1}(z_0)$. Then, $\mathcal{I}_\pi(z_0) \leq \big[\mathcal{I}_\theta^{-1}(z_0)\big]_{11}$. Therefore, the asymptotic variance of $\widehat{\pi}(z_0)$ is less than or equal to that of $\widetilde{\pi}(z_0)$. We now show the asymptotic bias of $\widehat{\pi}(z_0)$ is less than that of $\widetilde{\pi}(z_0)$. Let η_j denote the jth entry of $b(z_0)$.
Then,

$$\begin{aligned}
|\widehat{b}(z_0)| &= |\mathcal{I}_\pi^{-1}(z_0)\eta_1| \\
&= |\big[\mathcal{I}_\theta^{-1}(z_0)\big]_{11}\eta_1| \\
&\geq |\big[\mathcal{I}_\theta^{-1}(z_0)\big]_{11}\eta_1 + \boldsymbol{a}_1\mathcal{I}_\theta^{-1}\boldsymbol{a}_2| \\
&= |b_1(z_0)|,
\end{aligned}$$

where $\boldsymbol{a}_1 = (1, 0, \ldots 0)^{\mathrm{T}}$ and $\boldsymbol{a}_2 = (0, \eta_2, \ldots, \eta_{p+1})^{\mathrm{T}}$. Since $\boldsymbol{a}_1\mathcal{I}_\theta^{-1}\boldsymbol{a}_2 > 0$, it follows that $|\widehat{b}(z_0)| \leq |b_1(z_0)|$, and hence, the asymptotic bias of $\widehat{\pi}(z_0)$ is less or equal to that of $\widetilde{\pi}(z_0)$. □

Ascent Properties

Proof of 2.2.6. 1. Define the class membership indicator as

$$\mathcal{C}_i = \begin{cases} 0 & \text{if } Y_i \text{ from Poisson state} \\ 1 & \text{if } Y_i \text{ from degenerate state.} \end{cases}$$

Assume that the complete data are $\{(W_i, \boldsymbol{X}_i, Y_i, \mathcal{C}_i\}_{i=1}^n$ random samples from the population $(W, \boldsymbol{X}, Y, \mathcal{C})$. Then, the distribution of $\mathcal{C}|W, \boldsymbol{X}, Y$ is

$$\mathbb{P}\big(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}\big)=\begin{cases}\frac{\pi(w)}{\pi(w)+(1-\pi(w))\exp(-\mu)} & Y=0\\ 0 & Y>0\end{cases}. \tag{2.46}$$

Consequently,

$$\mathbb{P}\big(\mathcal{C}=0|W,\boldsymbol{X},Y;\boldsymbol{\theta}\big)=\begin{cases}\frac{(1-\pi(w))\exp(-\mu)}{\pi(w)+(1-\pi(w))\exp(-\mu)} & Y=0\\ 1 & Y>0\end{cases}. \tag{2.47}$$

Letting $\boldsymbol{\theta}^{(t)}(W_i)=\{\pi^{(t)}(W_i), \boldsymbol{\beta}^{(t)}(W_i)\}$, then $\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}(W_i))=r_i^{(t+1)}$. Then,

$$\begin{aligned}\ell_{z_0}^{S_1}(\boldsymbol{\theta})&=\sum_{i=1}^n\log\big[f(y_i|W_i,\boldsymbol{X}_i;\boldsymbol{\theta})\big]K_h(W_i-z_0)\\ &=\sum_{i=1}^n\log\big[f(y_i|W_i,\boldsymbol{X}_i;\boldsymbol{\theta})\big]\big(r_i^{(t+1)}+(1-r_i^{(t+1)})\big)K_h(W_i-z_0)\\ &=\sum_{i=1}^n\Bigg[\log\big[f(y_i|W_i,\boldsymbol{X}_i;\boldsymbol{\theta})\big]r_i^{(t+1)}K_h(W_i-z_0)\\ &+\log\big[f(y_i|W_i,\boldsymbol{X}_i;\boldsymbol{\theta})\big](1-r_i^{(t+1)})K_h(W_i-z_0)\Bigg].\end{aligned} \tag{2.48}$$

Note, that if $y_i=0$, then 2.46 and 2.47 imply

$$\log\big[f(0|W_i,\boldsymbol{X}_i;\boldsymbol{\theta})\big]=\begin{cases}\log\big[(1-\pi(w_i))\mu_i\big]-\log\Big[\mathbb{P}\big(\mathcal{C}=0|W,\boldsymbol{X},Y;\boldsymbol{\theta}\big)\Big] & C_i=0\\ \log(\pi(w_i))-\log\Big[\mathbb{P}\big(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}\big)\Big] & C_i=1.\end{cases} \tag{2.49}$$

Then, substituting 2.49 into 2.48, we obtain

$$\begin{aligned}\ell_{z_0}^{S_1}(\boldsymbol{\theta})&=\sum_{\{y_i=0\}}\Bigg[\Big(\log(\pi_i)-\log\big(\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i;\boldsymbol{\theta})\big)\Big)r_i^{(t+1)}\\ &+\Big(\log\big((1-\pi(w_i))\mu_i\big)-\log(\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i;\boldsymbol{\theta}))\Big)(1-r_i^{(t+1)})\Bigg]\times K_h(W_i-z_0)\\ &+\sum_{\{y_i>0\}}\log\big[(1-\pi(w_i))p(y_i|\boldsymbol{x}_i;\mu_i)\big](1-r_i^{(t+1)})K_h(W_i-z_0)\\ &=\sum_{i=1}^n r_i^{(t+1)}\log(\pi)K_h(W_i-z_0)+\sum_{i=1}^n(1-r_i^{(t+1)})\log\big[(1-\pi)p(y_i|\boldsymbol{x}_i;\mu_i)\big]K_h(W_i-z_0)\\ &-\sum_{i=1}^n\Big[\log(\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}))r_i^{(t+1)}+\log(\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}))(1-r_i^{(t+1)})\Big].\end{aligned} \tag{2.50}$$

Therefore,

$$
\begin{aligned}
n^{-1}\big[\ell_{z_0}^{S_1}(\boldsymbol{\theta}^{(t+1)}) - \ell_{z_0}^{S_1}(\boldsymbol{\theta}^{(t)})\big] &= n^{-1}\Bigg[\Big\{\ell_{z_0,C}^{S_1}(\boldsymbol{\theta}^{(t+1)}) - \sum_{i=1}^{n}\Big[\log(\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)}))r_i^{(t+1)} \\
&\quad + \log(\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)}))(1-r_i^{(t+1)})\Big\} \\
&\quad - \Big\{\ell_{z_0,C}^{S_1}(\boldsymbol{\theta}^{(t)}) - \sum_{i=1}^{n}\Big[\log(\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)}))r_i^{(t+1)} \\
&\quad + \log(\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)}))(1-r_i^{(t+1)})\Big\}\Bigg] \\
&= n^{-1}\big[\ell_{z_0,C}^{S_1}(\boldsymbol{\theta}^{(t+1)}) - \ell_{z_0,C}^{S_1}(\boldsymbol{\theta}^{(t)})\big] \\
&\quad - n^{-1}\sum_{i=1}^{n}\Bigg[\log\left(\frac{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}r_i^{(t+1)}\right) \\
&\quad + \log\left(\frac{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}(1-r_i^{(t+1)})\right)\Bigg] \times K_h(W_i - z_0).
\end{aligned}
$$

We know that $n^{-1}\big[\ell_{z_0,C}^{S_1}(\boldsymbol{\theta}^{(t+1)}) - \ell_{z_0,C}^{S_1}(\boldsymbol{\theta}^{(t)})\big] \geq 0$ based on the M-Step. Thus, we need to show in the second term

$$
\begin{aligned}
&- \liminf_{n\to\infty} n^{-1}\sum_{i=1}^{n}\Bigg[\log\left(\frac{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}r_i^{(t+1)}\right) \\
&+ \log\left(\frac{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}(1-r_i^{(t+1)})\right)\Bigg] \times K_h(W_i - z_0) \\
&= \limsup_{n\to\infty}\ - n^{-1}\sum_{i=1}^{n}\Bigg[\log\left(\frac{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}r_i^{(t+1)}\right) \\
&+ \log\left(\frac{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}(1-r_i^{(t+1)})\right)\Bigg] \times K_h(W_i - z_0) \\
&\geq 0
\end{aligned}
\tag{2.51}
$$

in probability. Equivalently, we will show that

$$
\begin{aligned}
\limsup_{n\to\infty}\ n^{-1}\sum_{i=1}^{n}&\Bigg[\log\left(\frac{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}r_i^{(t+1)}\right) \\
&+ \log\left(\frac{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}(1-r_i^{(t+1)})\right)\Bigg] \times K_h(W_i - z_0) \\
&\leq 0
\end{aligned}
$$

Note that if $Y_i > 0$, then $r_i^{(t+1)} = 0$ and $\log\big(\mathbb{P}(\mathcal{C}=1|W_i;\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})\big) = -\infty$. Thus, we will use the usual measure-theoretic convention that $0 \times \infty = 0$.

Define

$$L_g = n^{-1}\sum_{i=1}^{n}\Bigg[\log\left(\frac{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}r_i^{(t+1)}\right) + \log\left(\frac{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}(1-r_i^{(t+1)})\right)\Bigg]\times K_h(W_i-z_0),$$

and

$$L_J = n^{-1}\sum_{i=1}^{n}\log\Bigg[\left\{\frac{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=1|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}\right\}r_i^{(t+1)} + \left\{\frac{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}\right\}(1-r_i^{(t+1)})\Bigg]\times K_h(W_i-z_0).$$

Note, since $\log(\cdot)$ is convex, Jensen's inequality gives $L_g \leq L_J$. We now show $\limsup_{n\to\infty}\mathbb{P}(\|L_J\|^2\epsilon) = 0$ for any $\epsilon > 0$. For simplicity, when $Y_i = 0$ assume that $r_i^{(t)} \geq a > 0$ for small a. This can always be done in practice by taking the minimum of the $r_i^{(t+1)}$. Notice,

$$\mathbb{E}(L_J) = \mathbb{E}\Bigg(\log\Bigg[\left\{\frac{\mathbb{P}(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t)})}\right\}\mathbb{P}(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)}) + \left\{\frac{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}\right\}\mathbb{P}(\mathcal{C}=0|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)})\Bigg]\times K_h(W_i-z_0)\Bigg).$$

Define,

$$\Delta_n(\boldsymbol{X},Y) := \mathbb{E}\Bigg(\log\Bigg[\left\{\frac{\mathbb{P}(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t)})}\right\}\mathbb{P}(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)}) + \left\{\frac{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=0|W_i,\boldsymbol{X}_i,Y_i;\boldsymbol{\theta}^{(t)})}\right\}\mathbb{P}(\mathcal{C}=0|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)})\Bigg]\times K_h(W_i-z_0)|\boldsymbol{X},Y\Bigg).$$

Note that the random variable we are taking the expectation with respect to is dominated by the quantity $Q = \log\Big\{a^{-1}\big(\mathbb{P}(\mathcal{C}=1|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)}) + \mathbb{P}(\mathcal{C}=0|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)})\big)\Big\}\times K_h(Z_i-z_0)$, which has finite expectation. Moreover, when conditioning on $\boldsymbol{X}, Y$, for sufficiently large t,

$$\lim_{n\to\infty}\frac{\mathbb{P}(\mathcal{C}=c|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=c|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t)})} = 1$$

Therefore, by the Lebesgue Dominated Convergence Theorem,

$$\begin{aligned}\liminf_{n\to\infty}\Delta_n(\boldsymbol{X},Y) &= \mathbb{E}\Bigg(\liminf_{n\to\infty}\left\{\sum_{c=0}^{1}\frac{\mathbb{P}(\mathcal{C}=c|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=c|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t)})}\mathbb{P}(\mathcal{C}=c|W,\boldsymbol{X},Y;\boldsymbol{\theta}^{(t+1)})\right\}\Bigg)\\ &= \mathbb{E}\big(\log(1)\times K_h(W-z_0)|\boldsymbol{X},Y\big)\\ &= 0.\end{aligned}$$

Since $\Delta_n(\boldsymbol{X}, Y)$ is bounded, it follows by law of total expectation that

$$\liminf_{n\to\infty} \mathbb{E}(L_J) = \liminf_{n\to\infty} \mathbb{E}(\Delta_n(\boldsymbol{X}, Y)) = 0.$$

We now calculate the $\text{Var}(L_J)$. Notice that the $\text{Var}(L_J)$ is dominated by

$$\begin{aligned}
S &= n^{-1}\mathbb{E}\Bigg(\log\Bigg[\sum_{c=0}^{1} \frac{\mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t)})} \mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t+1)})\Bigg] K_h(W - z_0)\Bigg)^2 \\
&= n^{-1}\int \Bigg(\log\Bigg[\sum_{c=0}^{1} \frac{\mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t)})} \mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t+1)})\Bigg] K_h(W - z_0)\Bigg)^2 f(w)dw \\
&= n^{-1}\int \Bigg(\log\Bigg[\sum_{c=0}^{1} \frac{\mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t+1)})}{\mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t)})} \mathbb{P}(\mathcal{C}=c; W, \boldsymbol{X}, Y; \boldsymbol{\theta}^{(t+1)})\Bigg] W_h(z - z_0)\Bigg)^2 \\
&\Big(f(z_0) + f'(z_0)(w - z_0) + O(\|w - z_0\|^2)\Big)dw \\
&\leq a^{-1}(nh)^{-1}\big(f(z_0)\int h^{-1}K^2(h^{-1}(w - z_0))dw + f'(z_0)\int h^{-1}K^2(h^{-1}(w - z_0))(w - z_0) \\
&+ O(\|z - z_0\|^2)\big) \\
&= O((nh)^{-1}) \\
&= o(1).
\end{aligned}$$

Therefore, $\limsup_{n\to\infty} \mathbb{E}(\|\{L_J\|^2) = 0$; hence, $\limsup_{n\to\infty} L_J \leq 0$ in probability.
Thus, $\limsup_{n\to\infty} n^{-1}\big[\ell(\boldsymbol{\theta}^{t+1}(z_0)) - \ell(\boldsymbol{\theta}^{(t)}(z_0))\big] \geq 0$, which completes the proof.

2. The ascent property of ℓ_2 follows immediately from the ascent property of ordinary EM algorithms.
3. Notice that

$$\begin{aligned}
\ell_{z_0}^{S_3}(\pi^{(t+1)}) - \ell_{z_0}^{S_3}(\pi^{(t)}) &= \sum_{i=1}^{n} \log\Bigg\{ \frac{f(y_i|w_i, \boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}, \pi^{(t+1)})}{f(y_i|w_i, \boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}, \pi^{(t)})} \Bigg\} K_h(w_i - z_0) \\
&= \sum_{i=1}^{n} \log\Bigg\{ \Big[\frac{\pi^{(t)} I\{y_i = 0\}}{f(y_i|w_i, \boldsymbol{x}_i; \pi^{(t)}, \widehat{\boldsymbol{\beta}})}\Big]\Big[\frac{\pi^{(t+1)} I\{y_i = 0\}}{\pi^{(t)} I\{y_i = 0\}}\Big] \\
&+ \Big[\frac{(1 - \pi^{(t)})p(y_i; \widehat{\boldsymbol{\beta}})}{f(y_i|w_i, \boldsymbol{x}_i; \pi^{(t)}; \widehat{\boldsymbol{\beta}})}\Big]\Big[\frac{\pi^{(t+1)} p(y_i; \widehat{\boldsymbol{\beta}})}{\pi^{(t)} p(y_i; \widehat{\boldsymbol{\beta}})}\Big] \Bigg\} K_h(w_i - z_0) \\
&= \sum_{i=1}^{n} \log\Bigg\{ r_i^{(t+1)}\Big[\frac{\pi^{(t+1)} I\{y_i = 0\}}{\pi^{(t)} I\{y_i = 0\}}\Big] \\
&+ (1 - r_i^{(t+1)})\Big[\frac{\pi^{(t+1)} p(y_i; \widehat{\boldsymbol{\beta}})}{\pi^{(t)} p(y_i; \widehat{\boldsymbol{\beta}})}\Big] \Bigg\} K_h(W_i - z_0)
\end{aligned}$$

Based on Jensen's inequality, it follows

$$\begin{aligned}
\ell_{z_0}^{S_3}(\pi^{(t+1)}) - \ell_{z_0}^{S_3}(\pi^{(t)}) &\geq \sum_{i=1}^{n} \Bigg[r_i^{(t+1)} \log\Big[\frac{\pi^{(t+1)} I\{y_i = 0\}}{\pi^{(t)} I\{y_i = 0\}}\Big] + (1 - r_i^{(t+1)}) \log\Big[\frac{\pi^{(t+1)} p(y_i; \widehat{\boldsymbol{\beta}})}{\pi^{(t)} p(y_i; \widehat{\boldsymbol{\beta}})}\Big]\Bigg] \\
&= \ell_{z_0, C}^{S_3}(\pi^{(t+1)}, \widehat{\boldsymbol{\beta}}) - \ell_{z_0, C}^{S_3}(\pi^{(t)}, \widehat{\boldsymbol{\beta}}) \\
&\geq 0
\end{aligned}$$

Table 2.7: Examination of Average MSE and Average RASE.

	Bandwidth($n = 200$)				Bandwidth($n = 400$)			
MSE	1.07	2.18	4.36	PAR	1.07	2.18	4.36	PAR
β_0	0.016	0.0160	0.0202	0.3425	$2e^{-4}$	$1e^{-5}$	$3e^{-4}$	0.2052
β_1	0.0004	0.0004	0.0005	1.6191	$3e^{-5}$	$2e^{-5}$	$2e^{-6}$	1.6834
				RASE_π				
π	0.0727	0.08423	0.0898	0.0962	0.0508	0.0705	0.0768	0.0856

where the final inequality is based on the M-Step of the complete local likelihood.

□

Additional Simulation Study

Y is generated from the ZIP Regression model with a single covariate $X \sim \text{Unif}(0, 10)$, where X is applied in both the Poisson and zero-inflation state. The true $\beta = (.1, .2)^{\text{T}}$, and the zero-inflation probability has the form

$$\pi(x) = (3 + \sin(x))^{-1}.$$

Again, we study MSE of $\widehat{\boldsymbol{\beta}}$ and RASE of $\widehat{\pi}(\cdot)$ of the undersmoothed, CV, and oversmoothed bandwith, along with the parametric ZIP regression model. The numerical results can be seen in Table 2.7. We see that in general, the undersmoothed bandwidth performs the best for both estimation of $\boldsymbol{\beta}$ and π. The CV and oversmoothed bandwidth gives similar accuracy for estimation of $\boldsymbol{\beta}$, but the oversmoothed bandwidth does poorly at estimating π. Lastly, the parametric model is unsatisfactory at estimating both parameters.

We then examine coverage rates for the bootstrap Z, percentile, and BC intervals described previously. The Z-intervals coverage for $\boldsymbol{\beta}$ is given in Table 2.8. Overall, the coverage rates are under the nominal level. This could be a situation in which bootstrap calibration may be useful due to the high curvature of π, which leads to signifcantly underestimating the standard errors. In general, the three models have similar coverage across $n = 200$ and $n = 400$. As with the first simulation study, we do see a drop in coverage as n increases.

The estimation of π from a single Monte-Carlo replicate can be seen in Figure **??**. We see that the true function is quite varialbe across the range of 0-10, and thus it is unsuprising that it is difficult to estimate. From the figure, we see that the undersmoothed bandwidth does the best job of estimating π, and the parametric and oversmoothed models cannot correctly estimated the curvature of π. Moreover, there is critical points at $\pi/2$ and $5\pi/2$,

Table 2.8: Coverage Resuls $\boldsymbol{\beta}$.

Parameter	95%	95%
	$n = 200$,$h = 1.07$	$n = 400$,$h = 1.07$
β_0	89.80	84.80
β_1	91.20	89.20
	$n = 200$,$h = 2.18$	$n = 400$,$h = 2.18$
β_0	89.20	82.60
β_1	91.80	87.60
	$n = 200$,$h = 4.36$	$n = 400$,$h = 4.36$
β_0	82.60	78.80
β_1	90.00	86.00

and no estimates the critical points correctly. This is a case where a local polynomial fit would be useful, or possibly a locally adaptive bandwidth.

The coverage results for π can be seen in 2.9. Overall, we see again that the undersmoothed bandwidth with the BC intervals have the best coverage, which is the only interval that consisantly gets near nominal coverage, although no interval is desirable. The explanation for this is that $\pi(\cdot)$ has large total variation, especially compared to $\pi(\cdot)$ from the first simulation study.

Table 2.9: Coverage Rates for Intervals for π. The three numbers in each cell represent the coverage for the Z, percentile, and BC interval, respectively, for a value of x and h.

	$n = 200$			$n = 400$		
h	1.07	2.18	4.36	1.07	2.18	4.36
	92.40	80.00	58.60	89.40	59.40	30.60
1	92.40	67.00	52.20	82.40	35.20	24.60
	94.20	82.80	56.00	92.00	61.60	23.00
	89.40	65.40	51.20	79.80	38.20	25.60
2	81.60	57.80	49.00	60.00	27.80	21.60
	93.20	71.80	48.80	87.20	40.00	17.60
	90.20	86.20	85.40	85.60	77.00	78.80
3	91.80	87.00	78.6	83.00	74.80	72.80
	92.20	88.00	83.20	89.20	73.40	73.60
	88.60	57.20	25.00	85.50	34.40	5.80
4	83.60	44.80	22.60	73.00	19.20	4.40
	91.40	69.20	35.00	88.60	46.60	12.40
	74.40	19.40	5.00	60.80	3.40	0.20
5	57.80	10.60	3.60	32.80	0.20	0.00
	84.40	31.20	8.40	69.20	6.40	0.02
	93.20	89.60	80.40	94.60	90.40	73.80
6	94.40	86.20	76.00	95.00	82.20	70.60
	93.00	91.80	83.00	92.20	90.40	77.40
	91.60	71.80	65.40	86.20	48.60	43.80
7	86.00	69.60	62.00	78.80	46.40	37.60
	93.40	73.20	65.00	90.20	52.80	40.40
	92.80	68.60	46.60	82.20	41.60	20.40
8	87.00	60.00	48.60	78.40	28.80	22.20
	93.60	74.00	48.00	91.80	54.20	18.20
	96.20	93.60	81.80	93.00	92.00	73.40
9	97.20	91.80	76.40	94.80	95.80	59.00
	95.20	94.20	84.20	93.40	91.20	73.00

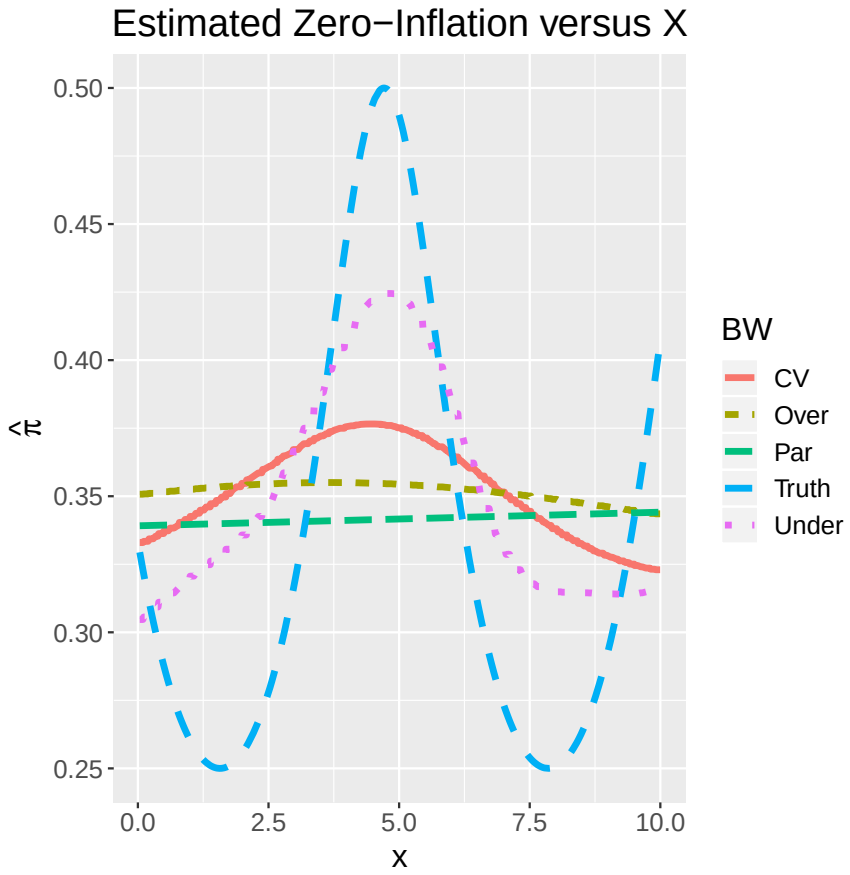


Figure 2.13: Estimated Zero-Inflation for Simulation Study 2

Chapter 3 Conclusions and Future Directions

3.1 Conclusions

In Chapter I, the utility of zero-inflated regression models was discussed, along with recent contributions to the literature. In Chapter 2, taking inspiration from the mixture-of-regressions literature, a novel semiparametric ZIP regression model was developed. The assumption of globally logit-linear zero-inflation probabilites was relaxed, and instead a non-parametric form for the mixing proportion was assumed. This weaker condition allows us flexibility in modeling the zero-inflation probabilies, and allows us to evaluate the assumption of logit linear mixing proportions in the parametric model. This could be useful because we do not observe the "switching-process" of the degenerative state and Poisson state, and the semiparametric model gives us a way to evaluate goodness of fit of the parametric switching process.

Since our model is similar to the partially linear model, a three step backfitting procedure was proposed for estimation of the parameters. The backfitting algorithm alternates between non-parametric and parametric estimation. A "EM like" algorithm was utilized at each estimation step. Asymptotic properties of the backfitting estimators was established, along with ascent properties of the objective functions at each estimation step.

Further, inference in the semiparametric framework was discussed. The asymptotic variance of both $\boldsymbol{\beta}$ and π was estimated via the bootstrap, and confidence intervals for $\pi(z)$ were constructed by three methods - Z, percentile, and BC intervals. From our simulations, the BC intervals for $\pi(z)$ were the most reliable in terms of coverage probability. For $\boldsymbol{\beta}$, the Z-intervals were found to have satisfactory coverage probabilites, although there seems to be an unexpected decrease in coverage as n increases. This may signify the standard errors are being underestimated for large n, and some calibration is needed.

After discussing confidence intervals, hypothesis testing in the semiparmetric setting via the bootstrap LRT was discussed. Our simulations indicate that the bootstrap LRT is a consistant test, along with preserving the nominal type-I error rate. Issues arose around negative bootstrap LRT statistics, but we anticipate that the issue can be mediated with either a bootstrap estimation of bias, or using an undersmoothed bandwidth for testing.

Finally, the proposed semiparametric ZIP model was utilized on an Alzheimer's and meth lab seizure data set. It was seen that the semiparametric regression model confirmed the reasonableness of the parametric ZIP regression model. Moreover, in practice, it is imperative to examine multiple bandwidths since undersmoothed and CV bandwidths can give spurious results, but the oversmoothed bandwidth can miss interesting features.

3.2 Future Directions

Semiparametric Regression Model

From the second simulation study in Section 2.6, it was seen that the bias of the local constant kernel estimator for $\pi(\cdot)$ is substantial when the curve has substantial total variation. We anticipate that this issue can be resolved by using local polynomial regression with degree 2 or 3. Therefore, even though the regression function in Section 2.6 is unlikely to be seen in practice, it may be desirable to fit a higher degree polynomial (locally) to reduce the bias. Moreover, polynomials of degree d allow us to estimate the $1^{\text{st}}, , \ldots, d^{\text{th}}$ derivative of $\pi(\cdot)$, which may be of importance in practice for estimating critical or inflection points. The tradeoff is the "EM-like" algorithm for estimation would be more computationally intesive; although, the *locfit* package in R [141] gives computationally efficient estimation via local likelihood. Another solution to the intense compuational time would be to apply the one-step local quasi-likelihood estimator of [161]. In this paper, the authors propose a one-step solution to local likelihood estimation to mitigate the compuational resources.

In addition to higher degree polynomial fits, it would also be desired to extend model 2.12 to the ZINB distribution. As noted before, the ZINB model allows for overdispersion, which is common in practice. We anticipate estimation of a semiparametric ZINB regression model would be analagous to the estimation steps discussed in Section 2.3, with the challange in estimation of the dispersion parameter, θ. A possible solution would be to combine the ECM algorithm for ZINB regression in Section 1.11, with the estimation steps in Section 2.3. Thus, a "ECM-like" algorithm could be developed. Whether the ascent properties and asymptotic normality of estimators follow would be of theoretical interest.

Related to a local ZINB regression model, it may be of use to model the dispersion parameter locally by a continuous covariate. The first property that would need to established is identifiability of the model, since the NB model is not in the GLM family when the dispersion parameter is unknown. Hence, the theorem for identifiability of mixtures-of-GLMs given in [140] would not apply.

Future Research Problems for Zero-Inflated Models

The seminal paper of [3] introduced the ZIP regression model and provided details about the likelihood estimation of the parameters. Many advancements with ZI count regression models have been made in the 26 years since the publication of that paper. We propose some interesting directions for research about ZI count regression models.

- [60] presented a novel measurement error (ME) regression model that accounts for both MEs in the covariates and zero-inflation for estimating the distributions of dietary intakes. The authors performed extensive empirical work and demonstrated

efficacy of this model in relating multiple dietary components and patterns with health outcomes. However, research needs to be performed in the context of ZI count regression models under different ME structures. For example, suppose that we are interested in relating our ZI count Y to a vector of covariaters, $\boldsymbol{X}$; however, $\boldsymbol{X}$ cannot be observed in practice. Instead, we observe $\boldsymbol{V}$. *Classical ME* is where $\boldsymbol{V} = \boldsymbol{X} + \boldsymbol{U}$ such that each component of $\boldsymbol{U}$, U_j for $j = 1, \ldots, p$, is $U_i \sim \mathcal{N}(0, \sigma^2)$. *Berkson ME* is where $\boldsymbol{X} = \boldsymbol{V} + \boldsymbol{U}$ (additive) or $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{U}$ (multiplicative), such that $\boldsymbol{V} \perp \boldsymbol{U}$, and $\mathbb{E}(\boldsymbol{U}) = \boldsymbol{0}$ or $\mathbb{E}(\boldsymbol{U}) = \boldsymbol{1}$, respectfully. Therefore, $\mathbb{E}(\boldsymbol{X}|\boldsymbol{V}) = \boldsymbol{V}$. Estimators and their properties for ZI count regression models need to be studied, as well as when ME occurs in the covariates for modeling the mixing proportion π. Some of the work of [162], who discussed ME in the context of non-ZI Poisson regression, could be leveraged for this research.

- Big data problems are of broad and current interest to researchers and data analysts. Zero-inflation can also occur in such big data problems, as highlighted with the census application [34]. One issue highlighted by the authors is the need for efficient computng routines when estimating ZI models applied to big data. In particular, routines are necessary to handle ultra-high dimensional variable selection in ZI count regression models. Such routines coculd be developed in the spirit of the *iteratively sure independent screening* approach of [163]. Perhaps even more beneficial will be including these computational routines in a statistical package devoted to modeling and inference tools for ZI count regression models. In Section 1.5, we highliighted major routines available in statistical software packages. However, most of these simply estimate ZIP and ZINB regression models, with options for obtaining simple residual summaries. A package that encompasses may of the modern methods that we discussed, including routines for big data problems, will make an invaluable contribution.

- Later in this chapter, we will note some Bayesian hierarchical models that have been developed for ZI counts in spatial data. One specific type of spatial data is *areal data*, which is aggregated quantities for each measured (areal) unit within some meaningful partition of a given region, such as counties within a state. A growing research topic is developing efficacious spatial regression models that caputure not only zero-inflation, but more generally characterize data dispersion for areal count data. Such models could better address problems related to the spread of diseases [164], trends in emergency department visits [99, 165], and changes in the status of housing units for conducting censuses [166]. One alternative to the models proposed for these applied problems is development of a spatial CMP regression model, which could provide a flexible framework for caputring the data dispersion.

- In Section 1.9, we discussed the notion of zero-inflation and diagonal-inflation in multivariate count regression models, with an emphasis on multivariate Poisson regression models. We noted some applied work where zero-inflation has been investigated for other multivariate count regression models. However, there is a need for a more rigorous development and treatment of ZI and DI count regression models beyond the multivariate Poisson regression setting. More generally, it would be beneficial to develop a unified framework about zero-inflation and diagonal-inflation in multivariate count regression models, regardless of the assumed count distribution. Such work could further inform more complex data structures, such as ZI counts in tensor regression. [167] developed an effective framework for tensor regression models that allows for discrete responses. However, the notion of zero-inflation has, to our knowledge, not been investigated.

- Variable selection in ZI regression models, and more generally mixtures-of-regressions models, is non-trivial task. This is due to the fact that covariates must be selected for both the count mean state and the zero-inflation state. Commonly, analysts rely on BIC to decide between competing models. Recently, [168] studied the *smoothly clipped absolute deviations* (SCAD) penalty for variable selection. To our knowledge, no other methodology has been studied for variable selection in ZI regression models.

3.3 Zero-Inflation in Spatial Data

CAR Regression Model

Most zero-inflated spatial regression models employ a conditionally autoregressive (CAR) covariance structure to model dependency. [46] was the first to utilize the CAR process to model abundance of isopod nest burrows. In similar work, [99] developed a Poisson hurdle model with a CAR prior to model ER visits in Durham County, NC. Lastly, [169] applied the CAR random effect model and a AR(1) time dependency to investigate harbor seal abundance. Due to the complexity of the likelihood function, the CAR ZI regression model is typically estimated via a Bayesian approach.

Before discussing the CAR model in the zero-inflation context, for illustration, we discuss the CAR model in the Gaussian data context. This is a summary of the discussion of [170]. Let $\boldsymbol{s} \in \mathbb{R}^2$, with $Y(\boldsymbol{s})$ denoting the outcome at at the grid location of $\boldsymbol{s}$. We can think of $Y(\boldsymbol{s}) = Y\{(s_1, s_2)\}$ as the outcome at a longitude of s_1 and latitude s_2. The CAR model assumes that

$$Y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j)\}_{j\neq i} = \mu_i + \rho \sum_{j=1}^{n} c_{ij}\big(y(\boldsymbol{s}_j) - \mu_j\big) + \epsilon_i, \tag{3.1}$$

where $\epsilon_i \sim \mathcal{N}(0, \tau^2)$, with

$$c_{ij} = \begin{cases} 0 & i = j \\ 1 & \text{sites i and j are spatial neighbors} \\ 0 & \text{otherwise .} \end{cases}$$

Here, ρ is a correlation parameter with $-1 < \rho < 1$. In interpretation, model 3.1 implies that the outcome at location $\boldsymbol{s}_i$ given the observed outcomes of $i \neq j$ has a site mean μ_i, and then the overall site mean is adjusted by the observed values, $y(\boldsymbol{s}_j)$, and site means, μ_j, for the neighboring locations $\boldsymbol{s}_j$ of $\boldsymbol{s}_i$. Here ρ controls the amount of correlation between the spatial neighbors $\boldsymbol{s}_j$ of $\boldsymbol{s}_i$. It can be shown using Brook's Lemma [171] that the joint distribution of

$$\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))^{\mathrm{T}} \sim \mathcal{N}_n\big(\boldsymbol{\mu}, \tau^2(I - \rho C)^{-1}\big),$$

where C is the $n \times n$ matrix with $C_{ij} = c_{ij}$.

The extension to the Spatial ZIP Regression model with CAR covariance sturcture is via mixed model. More formally,

$$\begin{aligned} \log(\mu(\boldsymbol{s}_i)) &= \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \delta_{1i}, \\ \operatorname{logit}(\pi(\boldsymbol{s}_i)) &= \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\alpha} + \delta_{2i}, \end{aligned} \tag{3.2}$$

where $\boldsymbol{\delta_k} \sim \mathcal{N}(\mathbf{0}, \tau_k^2(I - \rho_k C)^{-1})$ for $k = 1, 2$. Some authors do not include δ_{2i} [46] in the model. Meanwhile other authors assume that $\boldsymbol{\delta}_1 \perp \boldsymbol{\delta}_2$, while other authors [99] assume there may be dependence between $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$. Then, the likelihood then is

$$L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = \prod_{i=1}^{n} f(y_i | \delta_{1i}, \delta_{2i}; \boldsymbol{\theta}) g(\boldsymbol{\phi}_i), \tag{3.3}$$

where $f(\cdot)$ is the ZIP mass function and $g(\cdot)$ is the joint distribution of $\boldsymbol{\phi}_i = (\delta_{1i}, \delta_{2i})^{\mathrm{T}}$. The likelihood in 3.3 is difficult to maximize, and thus MCMC is typically utilized, with "uninformative" priors on the parameters.

Spatial Meth Data Analysis

It was noted in from the heat map in Section 2.6 that a spatial model could be of use for analyzing the meth lab seizure data set. Therefore, a Poisson negative binomial, and ZIP CAR regression model was fit. The ZIP CAR model was fit using the `CARBayes` [172] package in *R*. The covariates examined were percent poverty, the percent of rural area in a county, the amount of PSE sold in the county, and an indicator whether the county is off the I-65 or Cumberland Parkway (Off-High). Since most counties in Kentucky are rural, percent of rural area was made into an indicator of whether percent of rural area was bigger

Table 3.1: Summary of results for CAR model.

Model	β_0	**Off-High**(β_1)	**% Rural**(β_2)	**% Poor**(β_3)	α_0	PSE (α_1)	ρ
Poisson	-8.9866 (-9.1507,-8.8211)	1.6352 (1.3561,.9200)	*	*	*	*	0.4128 (0.1421,0.7457)
ZIP	-8.8322 (-9.0245,-8.6305)	0.4797 (0.1982,0.7972)	*	*	-3.6612 (-6.6794,-1.9799)	-0.7886 (-4.9693,2.1467)	.5838 (0.2845,0.8915)
NB	-9.2814	1.1350	0.6653	0.0357	*	*	0.7012
ZINB	-8.5879 (-9.0601,-5.9562)	0.4914 (-0.0131,0.9960)	0.6117 (0.1927,1.0308)	*	-4.514 (-7.4873,-1.541)	*	*

Table 3.2: Prediction accuracy on 2012 data set.

Model	**L**
Poisson	8.1254
NB	9.0765
ZIP	5.0663
ZINB	6.0049

tha 95%. Likewise, PSE sales was employed as indicator for PSE sales larger than 50 mg per 100 people. The priors were taken to be

$$\begin{aligned}
\boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, 100\boldsymbol{I}_{4\times 4}), \\
\boldsymbol{\alpha} &\sim \mathcal{N}(\mathbf{0}, 2), \\
\tau^2 &\sim \text{Inv-Gamma}(1, .001), \\
\rho^2 &\sim \text{Unif}(-1, 1),
\end{aligned}$$

where τ^2 and ρ are covariance parameters for the CAR model. The priors were taken to be more informative for $\boldsymbol{\alpha}$ since the MCMC sampler would not converge under larger variances. Different covariates were utilized in the models due to issues of convergence.

Moreover, the negative binomial CAR regression model was fit using the `copCAR` [173] package. Priors for $\boldsymbol{\beta}$ were similar to that of the ZIP regression model. The ZINB CAR model was not considered since we couldn't find any work in the spatial literature employing such a model, which could indicate non-identifiability or computational issues. However, a ZINB regression model was fit with area development district (ADD) as a random effect.

A summary of the model results can be seen Table 3.1. The inferences across regression parameters (that are similar.

For model comparison, the 2012 data set was utilized as a test set. Absolute loss, $L(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = n^{-1}\sum_{i=1}^{n}|y_i - \widehat{y}_i|$, was employed for determining prediction accuracy. A comparison of the aforementioned models can be seen in 3.2. Observe that the most accurate model is the ZIP spatial model.

County specific predictions for Fayette and Jefferson county for 2012 can be seen in Table 3.3. Observe that the Poisson, NB, and ZIP largely overpredict for Fayette county, while the ZINB is fairly accurate. Conversely, the ZINB massively underpredicts for Jefferson county, while the Poisson and NB largely overpredicts, but the ZIP is somewhat accurate.

Table 3.3: Predictions for Fayette and Jefferson County in 2012.

Model	**Poisson**	**NB**	**ZIP**	**ZINB**	**Observed**
Fayette	36.9975	52.8030	42.6333	3.5229	6
Jefferson	475.5564	387.3297	172.5647	4.9506	129

Bibliography

[1] A. C. Cohen, Jr. Estimating the Parameter in a Conditional Poisson Distribution. *Biometrics*, 16(2):203–211, 1960.

[2] J. Mullahy. Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, 33(3):341–365, 1986.

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

[4] T. G. Martin, B. A. Wintle, J. R. Rhodes, P. M. Kunhert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham. Zero Tolerance Ecology: Improving Ecological Inference by Modelling the Source of Zero Observations. *Ecology Letters*, 8(11):1235–1246, 2005.

[5] E. L. Boone, B. Stewart-Koster, and M. J. Kennard. A Hierarchical Zero-Inflated Poisson Regression Model for Stream Fish Distribution and Abundance. *Environmetrics*, 23(3):207–218, 2012.

[6] J. M. Potts and J. Elith. Comparing Species Abundance Models. *Ecological Modelling*, 199:153–163, 2006.

[7] A. F. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R.* Springer, New York, NY, 2009.

[8] G. Baetschmann and R. Winkelmann. Modeling zero-inflated count data when exposure varies: With an application to tumor counts. *Biometrical Journal*, 55(5): 679–686, 2013.

[9] K. C. H. Yip and K. K. W. Yau. On Modeling Claim Frequency Data in General Insurance with Extra Zeros. *Insurance: Mathematics and Economics*, 36(2):153–163, 2005.

[10] Jean-Philippe Boucher and Michel Denuit. Fixed versus random effects in Poisson regression models for claim counts: a case study with motor insurance. *Astin Bulletin*, 36(1):285–301, 2006.

[11] Y. Tang, L. Xiang, and Z. Zhu. Risk Factor Selection in Rate Making: EM Adaptive LASSO for Zero-Inflated Poisson Regression Models. *Risk Analysis*, 34(6):1112–1127, 2014.

[12] L. S. Sarul and S. Sahin. An Application of Claim Frequency Data Using Zero Inflated and Hurdle Models in General Insurance. *Journal of Business, Economics and Finance*, 4(4):732–743, 2015.

[13] S.-P. Miaou. The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions. *Accident Analysis and Prevention*, 26(4):471–482, 1994.

[14] J. A. List. Determinants of Securing Academic Interviews After Tenure Denial: Evidence from a Zero-Inflated Poisson Model. *Applied Economics*, 33(11):1423–1431, 2001.

[15] M-L. Sheu, T.-W. Hu, T. E. Keeler, M. Ong, and H.-Y. Sung. The Effect of a Major Cigarette Price Change on Smoking Behavior in California: A Zero-Inflated Negative Binomial Model. *Health Economics*, 13(8):781–791, 2004.

[16] J. M. Albert, W. Wang, and S. Nelson. Estimating Overall Exposure Effects for Zero-Inflated Regression Models with Application to Dental Caries. *Statistical Methods in Medical Research*, 23(3):257–278, 2014.

[17] D. S. Young. Mixtures of Regressions with Changepoints. *Statistics and Computing*, 24(2):265–281, 2014.

[18] K. K. W. Yau, K. Wang, and A. H. Lee. Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. *Biometrical Journal*, 45(4):437–452, 2003.

[19] R. Nishii and S. Tanaka. Modeling and Inference of Forest Coverage Ratio Using Zero-One Inflated Distributions with Spatial Dependence. *Enviornmental and Ecological Statistics*, 20(2):315–336, 2013.

[20] D. C. Heilbron. Zero-Altered and Other Regression Models for Count Data with Added Zeros. *Biometrical Journal*, 36(5):531–547, 1994.

[21] W. H. Greene. Fixed and Random Effects Models for Count Data. *SSRN eLibrary*, 2007.

[22] J. M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK, 2nd edition, 2011.

[23] M. Pandya, H. Pandya, and S. Pandya. Bayesian Inference on Mixture of Geometric with Degenerate Distribution: Zero Inflated Geometric Distribution. *International Journal of Research and Reviews in Applied Sciences*, 13(1):53–66, 2012.

[24] D. B. Hall. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*, 56(4):1030–1039, 2000.

[25] A. M. Garay, E. M. Hashimoto, E. M. M. Ortega, and V. H. Lachos. On Estimation and Influence Diagnostics for Zero-Inflated Negative Binomial Regression Models. *Computational Statistics and Data Analysis*, 55(3):1304–1318, 2011.

[26] D. B. Hall and Z. Zhang. Marginal Models for Zero Inflated Clustered Data. *Statistical Modelling*, 4(3):161–180, 2004.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[28] X.-L. Meng and D. B. Rubin. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, 80(2):267–278, 1993.

[29] J. van den Broek. A Score Test for Zero Inflation in a Poisson Distribution. *Biometrics*, 51(2):738–743, 1995.

[30] N. Janaskul and J. P. Hinde. Score Tests for Zero-Inflated Poisson Models. *Computational Statistics and Data Analysis*, 40(1):75–96, 2002.

[31] N. Janaskul and J. P. Hinde. Score Tests for Extra-Zero Models in Zero-Inflated Negative Binomial Models. *Communications in Statistics - Simulation and Computation*, 38(1):92–108, 2008.

[32] M. Ridout, J. Hinde, and C. G. B. Demétrio. A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. *Biometrics*, 57(1):219–223, 2001.

[33] D. Deng and S. R. Paul. Score Tests for Zero-Inflation and Over-Dispersion in Generalized Linear Models. *Statistica Sinica*, 15(1):257–276, 2005.

[34] D. S. Young, A. M. Raim, and N. R. Johnson. Zero-Inflated Modelling for Characterizing Coverage Errors of Extracts from the US Census Bureau's Master Address File. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1): 73–97, 2017.

[35] B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and the American Statistical Association, 1995.

[36] Q. H. Vuong. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333, 1989.

[37] P. Wilson. The Misuse of the Vuong Test for Non-Nested Models to Test for Zero-Inflation. *Economics Letters*, 127:51–53, 2015.

[38] M. Mittlböck and T. Waldhör. Adjustments for R^2-Measures for Poisson Regression Models. *Computational Statistics and Data Analysis*, 34(4):461–472, 2000.

[39] K. F. Sellers and A. Raim. A Flexible Zero-Inflated Model to Address Data Dispersion. *Computational Statistics and Data Analysis*, 99:68–80, 2016.

[40] P. K. Dunn and G. K. Smyth. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

[41] G. A. Dagne. Hierarchical Bayesian Analysis of Correlated Zero-Inflated Count Data. *Biometrical Journal*, 46(6):653–663, 2004.

[42] S. K. Ghosh, P. Mukhopadhyay, and J.-C. Lu. Bayesian Analysis of Zero-Inflated Regression Models. *Journal of Statistical Planning and Inference*, 136(4):1360–1375, 2006.

[43] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS - A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. *Statistics and Computing*, 10(4):325–337, 2000.

[44] H. Jang, S. Lee, and S. W. Kim. Bayesian Analysis for Zero-Inflated Regression Models with the Power Prior: Applications to Road Safety Countermeasures. *Accident Analysis and Prevention*, 42(2):540–547, 2010.

[45] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, Florida, 2013.

[46] D. K. Agarwal, A. E. Gelfand, and S. Citron-Pousty. Zero-Inflated Models with Application to Spatial Count Data. *Environmental and Ecological Statistics*, 9(4): 409–426, 2002.

[47] M. J. Bayarri, J. O. Berger, and G. S. Datta. Objective Bayes Testing of Poisson Versus Inflated Poisson Models. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3 of *B. Clarke and S. Ghosal (eds.) IMS Collections*, pages 105–121. Institute of Mathematical Statistics, Beachwood, OH, 2008.

[48] F.-C. Xie, J.-G. Lin, and B.-C. Wei. Bayesian Zero-Inflated Generalized Poisson Regression Model: Estimation and Case Influence Diagnotics. *Journal of Applied Statistics*, 41(6):1383–1392, 2014.

[49] G. D. C. Barriga and F. Louzada. The Zero-Inflated Conway-Maxwell-Poisson Distribution: Bayesian Inference, Regression Modeling and Influence Diagnostic. *Statistical Methodology*, 21:23–34, 2014.

[50] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[51] B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, FL, 2nd edition, 2008.

[52] A. E. Gelfand, D. Dey, and H. Chang. Model Determination Using Predictive Distributions with Implementation via Sampling-Based Methods. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), pages 147–167. Oxford University Press, Oxford, UK, 1992.

[53] N. Klein, T. Kneib, and S. Lang. Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and Overdispersed Count Data. *Journal of the American Statistical Association*, 110(509):405–419, 2015.

[54] Brian Neelon and Dongjun Chung. The lzip: A bayesian latent factor model for correlated zero-inflated counts. *Biometrics*, 73(1):185–196, 2017.

[55] S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, pages 3571–3594, 2010.

[56] Brian H Neelon, A James OMalley, and Sharon-Lise T Normand. A bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, 10(4):421–439, 2010. doi: 10.1177/1471082X0901000404. URL `https://doi.org/10.1177/1471082X0901000404`. PMID: 21339863.

[57] David J. Spiegelhalter. Bayesian graphical modelling: a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1):115–133, 1998.

[58] Souparno Ghosh, Alan E. Gelfand, Kai Zhu, and James S. Clark. the k-zig: Flexible moedling of zero-inflated counts. *Biometrics*, 68(3), September 2012.

[59] Shiferaw Gurmu and Getachew A. Dagne. Bayesian approach to zero-inflated bivariate ordered probit regression model, with an application to tobacco use. *Journal of Probability and Statistics*, 2012, 2012. doi: https://doi.org/10.1155/2012/617678.

[60] S. Zhang, D. Midthune, P. M. Guenther, S. M. Krebs-Smith, V. Kipnis, K. W. Dodd, D. W. Buckman, J. A. Tooze, L. Freedman, and R. J. Carroll. A New Multivariate Measurement Error Model with Zero-Inflated Dietary Data, and Its Application to Dietary Assessment. *The Annals of Applied Statistics*, 5(2B):1456–1487, 2011.

[61] Lluis Bermúdex and Dimitris Karlis. Bayesian multivariate poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, 48(2):226–236, March 2011.

[62] SAS Institute Inc. *SAS/STAT® 9.4 User's Guide*. SAS Institute Inc., Cary, NC, 2013.

[63] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

[64] A. Zeileis, C. Kleiber, and S. Jackman. Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8):1–25, 2008. http://www.jstatsoft.org/v27/i08/.

[65] T. W. Yee. *Vector Generalized Linear and Additive Models, With an Implementation in R*. Springer, New York, NY, 2015.

[66] Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Mchler, and Benjamin M. Bolker. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, 9(2):378–400, 2017. URL `https://journal.r-project.org/archive/2017/RJ-2017-066/index.html`.

[67] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

[68] X. Liu, B. Winter, L. Tang, B. Zhang, Z. Zhang, and H. Zhang. Simulating Comparisons of Different Computing Algorithms Fitting Zero-Inflated Poisson Models for Zero Abundant Counts. *Journal of Statistical Computation and Simulation (in press)*, 2017.

[69] R. A. Rigby and D. M. Stasinopoulos. Generalized Additive Models for Location, Scale and Shape (with Discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 54(3):507–554, 2005.

[70] L. K. Muthén and B. O. Muthén. *Mplus Users Guide, 7^{th} Edition*. Muthén and Muthén, Los Angeles, CA, 1998–2012.

[71] Stata Technical Support. *Stata Statistical Software: Release 14*. StataCorp LP, College Station, TX, 2015.

[72] NCSS, LLC. *NCSS 11 Statistical Software*. Kaysville, UT, 2016.

[73] Piet de Jong and Gillian Z. Heller. *Generalized Linear Models for Insurance Data*. International Series on Actuarial Science. Cambridge University Press, 2008. doi: 10.1017/CBO9780511755408.

[74] Alicja Wolny-Dominiak and Michal Trzesiok. *insuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Non-life Insurance.*, 2014. URL `https://CRAN.R-project.org/package=insuranceData`. R package version 1.0.

[75] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[76] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference*. Springer-Verlag New York, New York, NY, 2^{nd} edition, 2002.

[77] Martyn Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling, 2003.

[78] F. Famoye and K. P. Singh. Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. *Journal of Data Science*, 4:117–130, 2006.

[79] P. C. Consul and G. C. Jain. A Generalization of the Poisson Distribution. *Technometrics*, 15(4):791–799, 1973.

[80] P. C. Consul. *Generalized Poisson Distributions: Properties and Applications*. Marcel Dekker, New York, NY, 1989.

[81] F. Famoye. Restricted Generalized Poisson Regression Model. *Communications in Statistics - Theory and Methods*, 22(5):1335–1354, 1993.

[82] C. Czado and A. Min. Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in a Zero-Inflated Generalized Poisson Regression. Technical report, Discussion Paper No. 423, Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, 2005. http://nbn-resolving.de/urn:nbn:de:bvb:19-epub-1792-8.

[83] P. L. Gupta, R. C. Gupta, and R. C. Tripathi. Score Test for Zero Inflated Generalized Poisson Regression Model. *Communications in Statistics - Theory and Methods*, 33 (1):47–64, 2005.

[84] C. Czado, V. Erhardt, A. Min, and S. Wagner. Zero-Inflated Generalized Poisson Models with Regression Effects on the Mean, Dispersion and Zero-Inflation Level Applied to Patent Outsourcing Rates. *Statistical Modelling*, 7(2):125–153, 2007.

[85] Y. Cui and W. Yang. Zero-Inflated Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci Underyling Count Trait with Many Zeros. *Journal of Theoretical Biology*, 256(2):276–285, 2009.

[86] R. W. Conway and W. L. Maxwell. A Queuing Model with State Dependent Service Rates. *Journal of Industrial Engineering*, 12(2):132–136, 1962.

[87] T. Imoto. A Generalized ConwayMaxwellPoisson Distribution which Includes the Negative Binomial Distribution. *Applied Mathematics and Computation*, 247:824–834, 2014.

[88] K. F. Sellers and G. Shmueli. A Flexible Regression Model for Count Data. *The Annals of Applied Statistics*, 4(2):943–961, 2010.

[89] K. F. Sellers, T. Lotze, and A. Raim. *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*, 2017. URL `https://CRAN.R-project.org/package=COMPoissonReg`. R package version 0.4.0.

[90] J. Feng and Z. Zhu. Semiparametric Analysis of Longitudinal Zero-Inflated Count Data. *Journal of Multivariate Analysis*, 102(1):61–72, 2011.

[91] B. Jørgensen. Exponential Dispersion Models (with Discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 49(2):127–162, 1987.

[92] M. J. Dobbie and A. H. Welsh. Modelling Correlated Zero-Inflated Count Data. *Australian and New Zealand Journal of Statistics*, 43(4):431–444, 2001.

[93] K.-Y. Liang, S. L. Zeger, and B. Qaqish. Multivariate Regression Analyses for Categorical Data. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 54(1):3–40, 1992.

[94] S. Iddi and G. Molenberghs. A Marginalized Model for Zero-Inflated, Overdispersed and Correlated Count Data. *Electronic Journal of Applied Statistical Analysis*, 6(2): 149–165, 2013.

[95] K. Wang, K. K. W. Yau, and A. H. Lee. A Zero-Inflated Poisson Mixed Model to Analyze Diagnosis Related Groups with Majority of Same-Day Hospital Stays. *Computer Methods and Programs in Biomedicine*, 68(3):195–203, 2002.

[96] Y. Min and A. Agresti. Random Effect Models for Repeated Measures of Zero-Inflated Count Data. *Statistical Modelling*, 5(1):1–19, 2005.

[97] K. F. Lam, H. Xue, and Y. B. Cheung. Semiparametric Analysis of Zero-Inflated Count Data. *Biometrics*, 62(4):996–1003, 2006.

[98] M. Alfò and A. Maruotti. Two-Part Regression Models for Longitudinal Zero-Inflated Count Data. *The Canadian Journal of Statistics*, 38(2):197–216, 2010.

[99] B. Neelon, P. Ghosh, and P. F. Loebs. A Spatial Poisson Hurdle Model for Exploring Geographic Variation in Emergency Department Visits. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 176(2):389–413, 2013.

[100] M. Yang, G. K. D. Zamba, and J. E. Cavanaugh. Markov Regression Models for Count Time Series with Excess Zeros: A Partial Likelihood Approach. *Statistical Methodology*, 14:26–38, 2013.

[101] F. Zhu. Zero-Inflated Poisson and Negative Binomial Integer-Valued GARCH Models. *Journal of Statistical Planning and Inference*, 142(4):826–839, 2012.

[102] E. Gonçalves, N. Mendes-Lopes, and F. Silva. Zero-Inflated Compound Poisson Distributions in Integer-Valued GARCH Models. *Statistics: A Journal of Theoretical and Applied Statistics*, 50(3):558–578, 2016.

[103] A. Arab, S. H. Holan, C. K. Wikle, and M. L. Wildhaber. Semiparametric Bivariate Zero-Inflated Poisson Models with Application to Studies of Abundance for Multiple Species. *Environmetrics*, 23(2):183–196, 2012.

[104] P. Faroughi and N. Ismail. Bivariate Zero-Inflated Generalized Poisson Regression Model with Flexible Covariance. *Communications in Statistics - Theory and Methods*, 46(15):7769–7785, 2017.

[105] D. Karlis and I. Ntzoufras. Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R. *Journal of Statistical Software*, 14(10):1–36, 2005.

[106] M. K. Olsen and J. L. Schafer. A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association*, 96(454):730–745, 2001.

[107] E. D. Mills. *Adjusting for Covariates in Zero-Inflated Gamma and Zero-Inflated Log-Normal Models for Semicontinuous Data*. PhD thesis, University of Iowa, 2013.

[108] Peter K. Dunn and Gordon K. Smyth. Evaluation of tweedie exponential dispersion model densities by fourier inversion. *Statistics and Computing*, 18(1):73–86, 2008.

[109] Gordon K. Smyth and Bent Jrgensen. Fitting tweedie's compound poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin*, 32(1):143157, 2002. doi: 10.2143/AST.32.1.1020.

[110] Masud Hasan and Peter K. Dunn. A simple poisson-gamma model for modelling rainfall occurance and amount simultaneously. *Agricultural and Forest Meteorology*, 150(10):1319–1330, 2010.

[111] Peter K. Dunn. *Tweedie: Evaluation of Tweedie Exponential Family Models*, 2017. R package version 2.3.0.

[112] R. Ospina and S. L. P. Ferrari. A General Class of Zero-or-One Inflated Beta Regression Models. *Computational Statistics and Data Analysis*, 56(6):1609–1623, 2012.

[113] J. Wieczorek and S. Hawala. A Bayesian Zero-One Inflated Beta Model for Estimating Poverty in U.S. Counties. In *JSM Proceedings, Section on Survey Research Methods*, pages 2812–2822, Alexandria, VA, 2011. American Statistical Association.

[114] H. Liu, S. Ma, R. Kronmal, and K.-S. Chan. Semiparametric Zero-Inflated Modeling in Multi-Ethnic Study of Atherosclerosis (MESA). *The Annals of Applied Statistics*, 6(3):1236–1255, 2012.

[115] Seong-Keon Lee and Seohoon Jin. Decision tree approaches for zero-inflated count data. *Journal of Applied Statistics*, 33(8):853–865, 2006. doi: 10.1080/02664760600743613. URL `https://doi.org/10.1080/02664760600743613`.

[116] W. R. Cupach and B. H. Spitzberg. *The Dark Side of Relationship Pursuit. From Attraction to Obsession and Stalking*. Lawrence Erlbaum Associates, Mahwah, NJ, 2004.

[117] T. Loeys, B. Moerkerke, O. De Smet, and A. Buysse. The Analysis of Zero-Inflated Count Data: Beyond Zero-Inflated Poisson Regression. *British Journal of Mathematical and Statistical Psychology*, 65(1):163–180, 2012.

[118] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 39(1):1–38, 1977.

[119] X.-L. Meng and D. B. Rubin. Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework. *Biometrika*, 80(2):267–278, 1993.

[120] H. Dai, Y. Bao, and M. Bao. Maximum Likelihood Estimate for the Dispersion Parameter of the Negative Binomial Distribution. *Statistics and Probability Letters*, 83(1):21–27, 2013.

[121] D. R. Hunter and D. S. Young. Semiparametric Mixtures of Regressions. *Journal of Nonparametric Statistics*, 24(1):19–38, 2012.

[122] Mian Huang and Weixin Yao. Mixture of regression models with varying mixing proportions: A semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724, 2012. doi: 10.1080/01621459.2012.682541. URL `https://doi.org/10.1080/01621459.2012.682541`.

[123] Richard E. Quandt. A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67(338):306–310, 1972. doi: 10.1080/01621459.1972.10482378. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10482378`.

[124] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2):181–214, 1994.

[125] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984. ISBN 0-534-98053-8.

[126] T. Benaglia, D. Chauveau, D. R. Hunter, and D. S. Young. mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, 32(6):1–29, 2009. http://www.jstatsoft.org/v32/i06/.

[127] Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics*, 29(1):153–193, 2001.

[128] J. Cao and W. Yao. Semiparametric mixture of binomial regressions with a degenerate component. *Statistica Sinica*, 22(1):27–46, 2012.

[129] Mian Huang, Runze Li, and Shaoli Wang. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108(503):929–941, 2013.

[130] Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414, 1982.

[131] Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, 22(2):580–615, 1994.

[132] X. He, H. Xue, and N.-Z. Shi. Sieve Maximum Likelihood Estimation for Doubly Semiparametric Zero-Infalted Poisson Models. *Journal of Multivariate Analysis*, 101 (9):2026–2038, 2010.

[133] H. Liu and K.-S. Chan. Generalized Additive Models for Zero-Inflated Data with Partial Constraints. *Scandinavian Journal of Statistics*, 38(4):650–665, 2011.

[134] Hai Liu and Kung sik Chan. Introducing cozigam: An r package for unconstrained and constrained zero-inflated generalized additive model analysis, 2011.

[135] D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric Regression*. Cambridge University Press, Edinburgh, Cambridge, 2003.

[136] R. Bellman. *Adaptive control processes*. Princeton University Press, Princeton, NJ, 1961.

[137] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York City, New York, 1985.

[138] C. Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 200.

[139] C.S. Li. Identifiability of zero-inflated poisson models. *Brazilian Journal of Probability and Statistics*, 23(3):306–312, 2012.

[140] S. Wang, W. Yao, and M. Huang. A Note on the Identifiability of Nonparametric and Semiparametric Mixtures of GLMs. *Statistics and Probability Letters*, 93:41–45, 2014.

[141] C. Loader. *Local regression and likelihood*. Springer-Verlag New York, New York City, New York, 1999.

[142] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer-Verlag Berlin Heidelberg, Berlin, Germany, 2004.

[143] Jianqing Fan, Nancy E. Heckman, and M. P. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150, 1995. doi: 10.1080/01621459.1995.10476496. URL `https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476496`.

[144] Runze Li and Hua Liang. Variable selection in semiparametric regression modeling. *The Annals of Statistics*, 36(1):261–286, 2008.

[145] Clive R. Loader. Bandwidth selection: classical or plug-in? *The Annals of Statistics*, 27(2):415–438, 1999.

[146] Tony Linderberg. *Scale-Space Theory in Computer Vision*. Springer US, 1994.

[147] Bart M. Haar Romeny. *Front-End Vision and Multi-Scale Image Analysis*. Springer Netherlands, 2003.

[148] Runze Li and J.S. Marron. Local likelihood sizer map. *The Indian Journal of Statistics*, 67(3):476–498, 2005.

[149] Samory Kpotufe and Vikas K Garg. Adaptivity to local smoothness and dimension in kernel regression. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3075–3083, USA, 2013. Curran Associates Inc. URL `http://dl.acm.org/citation.cfm?id=2999792.2999955`.

[150] Leo Breiman and Jerome Friedman. Estimating optimal transformations for multiple regressions and correlations (with discussion). *Journal of the American Statistical Association*, 80(391):580–619, 1985.

[151] Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989.

[152] Wolfgang Härdle and Adrian W. Bowman. Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of the American Statistical Association*, 83(401):102–110, 1988.

[153] Wolfgang Härdle and J.S. Marron. Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, 19(2):778–796, 1991.

[154] Michael H. Neumann and Jörg Polzehl. Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9(4):307–333, 1998.

[155] Bradley Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, NY, 1994.

[156] Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and the wilks phenomenon. *The Annals of Statistics*, 29(1):153–193, 2001.

[157] Jianqing Fan and Jiancheng Jiang. Nonparametric inference with generalized likelihood ratio tests. *TEST*, 16(3):409–444, 2007.

[158] M.A. Martin. On the double bootstrap. Technical Report 347, Department of Statistics, Stanford University, 1990.

[159] W.Y. Loh. Calibrating confidence coefficents. *Journal of the American Statistical Association*, 82(397):155–162, 1987.

[160] Mian Huang. *Nonparametric Techniques in Mixture of Regression Models*. PhD thesis, The Pennsylvania State University, 2009.

[161] Jianqing Fan and Jianwei Chen. One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Socity. Series B*, 62(4):927–943, 1999.

[162] J. Q. Guo and T. Li. Poisson Regression Models with Errors-in-Variables: Implications and Treatment. *Journal of Statistical Planning and Inference*, 104(2):391–401, 2002.

[163] J. Fan, R. Samworth, and Y. Wu. Ultrahigh Dimensional Feature Selection: Beyond The Linear Model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

[164] S. Gschlößl and C. Czado. Modelling Count Data with Overdispersion and Spatial Effects. *Statistical Papers*, 49(3):531–552, 2008.

[165] B. Neelon, H. H. Chang, Q. Liang, and N. S. Hastings. Spatiotemporal Hurdle Models for Zero-Inflated Count Data: Exploring Trends in Emergency Department Visits. *Statistical Methods in Medical Research*, 25(6):2558–2576, 2016.

[166] D. Musgrove, D. S. Young, J. Hughes, and L. E. Eberly. A Sparse Areal Mixed Model for Multivariate Outcomes, with an Application to Zero-Inflated Census Data. *Submitted*, 2017.

[167] H. Zhou, L. Li, and H. Zhu. Tensor Regression with Applications to Neuroimaging Data Analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

[168] Anne Buu, Norman Johnson, Runze Li, and Xianming Tan. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*, 30(18):2326–2340, 2011.

[169] Jay M. Ver Hoef and John K. Jansen. Space-time zero-inflated count models of harbor seals. *Environmetrics*, 18(7), 2007.

[170] N.A. Cressie. *Statistics for Spatial Data*. Wiley and Sons,Inc., Hoboken, New Jersey, 1993.

[171] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[172] Duncan Lee. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24, 2013. URL `http://www.jstatsoft.org/v55/i13/`.

[173] Emily Goren and John Hughes. *copCAR: Fitting the copCAR regression model for discrete areal data*. Denver, CO, 2017. R package version 2.0-2.

Vita

Education

- **University of Kentucky** — Lexington, Kentucky
  - *PhD in Statistics* — Aug. 2016 – May 2019
  - *Masters in Statistics* — Aug. 2014 – May 2016
- **University of Dayton** — Dayton, OH
  - *B.S. in Mathematics* — Aug. 2010 – May 2014

Work Experience

- **The Travelers Companies, Inc.** — Hartford,CT
  - *PI R&D Intern* — June 2018 – Aug. 2018
- **University of Kentucky** — Lexington, KY
  - *Research Assistant in the Applied Statistics Lab* — Aug. 2015 – present
  - *Teaching Assistant for STA 210* — Aug. 2014 – May 2015

Publications

- J. Evans., S. Wang, C. Greb, V. Kostas, C. Knapp, Q. Zhang, E. Roemmele, M. Stenger, and D. Randall. "Body Size Predicts Cardiac and Vascular Resistance Effects on Men's and Women's Blood Pressure". *Fronteirs in Physiology*, 8: 561, 2017.
- T. Rounsaville, C. Baskin, E. Roemmele, and M. Arthur. "Seed dispersal and site characteristics influence germination and seedling survival of the invasive liana *Euonymus fortunei* (wintercreepers) in a rural woodland". *Canadian Journal of Forest Research*, 48(11): 1343–1350, 2018.
- H. Seyyedhasani, J. Dvorak, and E. Roemmele. "Routing algorithm selection for field coverage planning based on field shape and fleet size". *Computers and Electronics in Agriculture*, 156 : 523–529, 2019.