2021

Novel Methods for Characterizing Conditional Quantiles in Zero-Inflated Count Regression Models

Xuan Shi
University of Kentucky, shixuan0817@gmail.com

Digital Object Identifier: https://doi.org/10.13023/etd.2021.188

Recommended Citation

Shi, Xuan, "Novel Methods for Characterizing Conditional Quantiles in Zero-Inflated Count Regression Models" (2021). *Theses and Dissertations--Statistics*. 56.
https://uknowledge.uky.edu/statistics_etds/56

Xuan Shi, Student

Dr. Derek S. Young, Major Professor

Dr. Katherine Thompson, Director of Graduate Studies

Novel Methods for Characterizing Conditional Quantiles in Zero-Inflated Count Regression Models

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
Xuan Shi

Lexington, Kentucky

Director: Dr. Derek Young, Professor of Statistics

Lexington, Kentucky

2021

ABSTRACT OF DISSERTATION

Novel Methods for Characterizing Conditional Quantiles in Zero-Inflated Count Regression Models

Despite its popularity in diverse disciplines, quantile regression methods are primarily designed for the continuous response setting and cannot be directly applied to the discrete (or count) response setting. There can also be challenges when modeling count responses, such as the presence of excess zero counts, formally known as zero-inflation. To address the aforementioned challenges, we propose a comprehensive model-aware strategy that synthesizes quantile regression methods with estimation of zero-inflated count regression models. Various competing computational routines are examined, while residual analysis and model selection procedures are included to validate our method. The performance of these methods is characterized through extensive Monte Carlo simulations. An application to the Oregon Health Insurance Experiment will also be discussed. We then extend our methods to the setting of longitudinal data with zero-inflated count responses, where the goal is to study identification and estimation of conditional quantile functions for such data. We first show that conditional quantile functions for discrete responses are identified in zero-inflated models with subject heterogeneity. Then, we develop a simple three-step approach to estimate the effects of covariates on the quantiles of the response variable. We present a simulation study to show the small sample performance of the estimator. Finally, we illustrate our model using the RAND Health Insurance Experiment data and the Combined Pharmacotherapies and Behavioral Interventions (COMBINE) data.

KEYWORDS: Zero-Inflated Model, Quantile regression, Discrete response, Continuous generalization, Nonlinear Least Squares, Subject Heterogeneity

Xuan Shi

May 14, 2021

Date

Novel Methods for Characterizing Conditional Quantiles in Zero-Inflated Count Regression Models

By

Xuan Shi

Derek Young

Director of Dissertation

Katherine Thompson

Director of Graduate Studies

May 14, 2021

Date

Dedicated to my family

ACKNOWLEDGMENTS

I would like to start by thanking my advisor, Dr. Derek Young, for his guidance and help on my research. Furthermore, thank you for the two valuable courses you taught in previous semesters. I would also like to thank my other committee members Dr. Carlos Lamarche, Dr. Solomon Harrar, Dr. Arnold Stromberg, Dr. Chenglong Ye and Dr. Xiangrong Yin for their time and suggestions. Moreover, thanks to the Univeristy of Kentucky Statistics Department for an unforgettable five years.

Next, I would like to thank my family for all the love and support from the very first day I was born. My parents have always been my role models. My father, born in a rural area in central China, works hard his whole life to be the best electrical engineer in the nation and brings a highly admired life to my family. My mother, once a respected accountant, gave up her professional career to look after me since middle school. My life would have been completely different without their working ethics and love. I would like to thank Zi Ye, my fiancee and future Mrs. Shi, for her patience and love.

Finally, I would like to thank the Ohio State University Statistics Department for preparing me for the academic maturity. Furthermore, I would like to thank my hometown, Huangshi, which means Yellowstone in Chinese. I would also like to thank my friends for all the happy memories we had together. Lastly, I would like to thank Guti, Raul Gonzalez, Ronaldo the Phenomenon, Real Madrid Football Club and John Ronald Reuel Tolkien.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Zero-Inflated Model

In practice, researchers often need to model data where the response variables are integer-valued. Discrete responses are discrete are used to represent counts of interest; for example, the counts of failures during a manufacturing process (Lambert, 1992), the number of motor vehicle crashes (Lord et al., 2005), or the number of housing units in a census block (Young et al., 2017).

One problem that frequently arises in analyzing count data is the presence of excess zeros. Count data with excess zeros are commonly found in many areas, including industry (Lambert, 1992), epidemiology (Bohning et al., 1999), ecology (Agarwal et al., 2002), transportation (Lord et al., 2005) and insurance (Baetschmann and Winkelmann, 2012). This phenomenon is formally known as *zero inflation*. When the data contains a higher proportion of zero counts than a model explains, this leads to biased estimation and misleading inference. To alleviate the effects of excess zeros, one effective class of models is zero-inflated (ZI) models. These models are parameterized as a mixture of two components. One component is the process of interest by the researchers. This discrete, conditional distribution determines the relationship between the response variable and the independent variables. The other component is a degenerate distribution at zero. This is used to account for the excess zeros in the observed dataset.

Lambert (1992) provided the fundamental paper on zero-inflated Poisson (ZIP) regression model, where the data is modeled as a mixture of a Poisson distribution and a degenerate distribution, and the model parameters are modeled as functions of covariates. The first component is considered as an imperfect state, where random zero counts occur by the Poisson distribution; the second component is considered as a perfect state, where a structural zero is the only possibility.

It is well-known that the expectation and the variance of a Poisson distribution are equal; hence, the ZIP model is most efficient when the variability matches the expected value. This, however, is not always a tenable assumption. Another ZI model, zero-inflated negative binomial (ZINB) model, is popular when the data shows additional overdispersion other than excess zeros. The negative binomial distribution has separate mean and dispersion parameters, hence, ZINB is more flexible in situations when the variability is greater than expected. Following the modeling strategies in Lambert (1992), Greene (1994) considered the modification of the negative binomial distribution to accomodate zero, and discussed the distinction between zero inflation and over-dispersion. Ridout et al. (2001) discussed ZINB models with inferences based on Score tests. Other works on ZINB models include Yau et al. (2003) and Mwalili et al. (2008).

Maximum likelihood estimation (MLE) for ZI modeling is commonly conducted via an expectation-maximization (EM) algorithm (Dempster et al., 1977) as in the original work by Lambert (1992). Another option is to use a Newton-Raphson algorithm, which has faster convergence than the EM algorithm; however, Newton-Raphson algorithm could fail to converge, as was noted by Lambert (1992).

Inferential aspects about ZI models have also been studied extensively in the literature. The most fundamental question is whether ZI count regression shows improvements over the corresponding count regression without ZI adjustments. This is equivalent to testing the presence of ZI. To conduct such tests, researchers utilize a score test (van den Broek, 1995; Jansakul and Hinde, 2002, 2008) or a boundary likelihood ratio test (Hilbe, 2011). Another popular choice is the Vuong's non-nested test based on likelihood ratio (Vuong, 1989).

Diagnostics based on residual analysis also provide straightforward yet insightful information in ZI count regression. Traditionally, researchers assess Pearson or deviance residuals given by the models. More insightful for count regression models are the randomized quantile residuals proposed by Dunn and Smyth (1996), which can be used for assessing model fitting in ZI count regression. For example, Young et al. (2017) utilized randomized quantile residuals in analyzing the quality of census frames for the 2020 Census.

The popularity of count regression models and ZI models is highlighted by its applications in various disciplines (Lambert, 1992; Bohning et al., 1999; Agarwal et al., 2002; Lord et al., 2005), however, most methods in this area focus solely on the mean structure of the conditional distribution for the response variable, given the independent variables. To better understand the relationship between the response variable and the independent variables, a natural extension of the analysis is to explore other aspects of the conditional distribution.

1.2 Quantile and Quantile Regression Model

Let Y be a random variable with cumulative distribution function (CDF) $F_Y(y) = P(Y \leq y)$. The τ^{th} quantile is defined as:

$$Q_Y(\tau) = \inf\left\{y : F_Y(y) \geq \tau\right\}, \tag{1.1}$$

where $0 < \tau < 1$ is the quantile level. Quantiles provide important information in characterizing a distribution, just like the expectation. However, the expectation summarizes the central tendency, while the quantiles can describe the complete distribution. Thus, the utilization of quantiles is preferred if the goal is to explore aspects other than the average.

When Y is a continuous random variable with strictly-increasing CDF F_Y, then,

$$Q_Y(\tau) = F_Y^{-1}, \tag{1.2}$$

where $Q_Y(\tau)$ is a strictly-increasing function of τ. Hence,

$$F_Y[Q_Y(\tau)] = P[Y \leq Q_Y(\tau)] = \tau \tag{1.3}$$

Of all the possible choices, perhaps the most famous quantile is obtained at $\tau = 0.5$, such that the resulting 50^{th} quantile is the median. In fact, for certain distributions (e.g., the normal distribution), the median equals the expectation. Hence, the quantile can be considered as an extension of the expectation.

Analysis of the quantiles, especially the median and the quartiles, has a long history (Galton, 1883), but one of the most significant advancements was due to Koenker and Bassett (1978) for developing linear quantile regression (QR). As in the case of classic linear regression, the goal of linear QR is to explore the conditional quantiles of the response, Y, given values of the independent variables, $\mathbf{X}$. Specifically,

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta}(\tau)\,, \tag{1.4}$$

where the regression coefficients, $\boldsymbol{\beta}(\tau)$ depends on the quantile level, τ. It is immediate from this notation, that the quantile regression coefficients can capture different effects given different quantiles.

The goal of linear QR is to explore the conditional quantiles of the response, Y, given values of the independent variables, $\mathbf{X}$. Since the seminal work by Koenker and Bassett (1978), countless theoretical results and methodological advancements have been made in this area. Numerous real data analyses have also been informed by QR. For a detailed discussion of QR literature and recent advancements, see Yu et al. (2003), Koenker (2005) and Koenker et al. (2017).

Computationally, there is usually no closed-form solution to the objective function in QR modeling, so results are obtained by numerical algorithms. Koenker and Bassett (1978) showed that the estimation can be transformed into a linear programming (LP) problem, where multiple algorithms are available. To date, the most popular choice is the simplex algorithm (Barrodale and Roberts, 1978). When the sample size is moderate, the Simplex algorithm is efficient and fast. The interior point algorithms are preferred for large sample sizes (Portnoy and Koenker, 1997) or nonlinear modeling (Koenker and Park, 1996). For more recent advancements in computational resources, see Geraci (2014, 2016).

The classic QR model does not assume a particular distribution for the response, and is mainly studied in a nonparametric framework; see Koenker and Bassett (1978), Takeuchi et al. (2006) and Chaudhuri (1991). More recently, parametric QR has been studied from a Bayesian perspective (Yue and Rue, 2011), where one of the most important advancements is the introduction of the asymmetric Laplace distribution (ALD) by Yu and Moyeed (2001). In a Bayesian setup, the use of the ALD is critical since a working likelihood is required. For a frequentist procedure, the ALD also provides a useful tool for likelihood-based inference, as in Geraci and Bottai (2007).

1.3 Quantile Regression Model for Discrete/ZI Data

When Y is a discrete random variable, there is no one-to-one relationship between $Q_Y(\tau)$ and τ since both the CDF and quantile functions are step functions. Furthermore, the non-differentiability inhibits the extension of the optimization routine as in the continuous case. Lastly, the linearity assumption does not hold for most problems when the response is a count.

A popular approach proposed by Machado and Santos Silva (2005) is to construct a continuous variable by jittering on the original counts. For the computational routine to work, an independent random variable that follows a continuous uniform distribution in $[0, 1)$, is generated. The random noise is then added to the original count response to transform it into a continuous variable. The traditional QR methods can then be applied to the updated data.

The main advantage of the jittering-based approach is that all of the existing QR methods can be applied easily after the transformation. However, this method does not guarantee the conditional quantiles for the new response to be the same as the conditional quantiles for the original response. Another concern with the jittering-based approach is quantiles crossing. By definition, the conditional quantile function $Q_Y(\tau|\mathbf{x})$ is a nondecreasing function of τ for any $\mathbf{x}$. This implies that, for $\tau_1 > \tau_2$,

$$Q_Y(\tau_1|\mathbf{x}) > Q_Y(\tau_2|\mathbf{x}). \tag{1.5}$$

Hence, quantiles crossing should be considered as an indicator of inaccurate estimation. In practice, however, the jittering-based procedure tends to incur quantiles crossing. This problem originates from the estimation routine in the traditional QR; furthermore, the random terms added to the count worsens the situation.

An alternative approach for QR with count data is the asymmetric maximum likelihood (AML) estimator proposed by Efron (1992). This estimator arises from the optimization of a smoothed objective function as given in Koenker and Bassett (1978) and hence is straightforward to interpret. However, the computation routine requires the quantile level, τ, to be greater than the proportion of zeros in the data. Thus, their method does not work for ZI settings. Newey and Powell (1987) proposed an asymmetric least squares (ALS) estimation that is akin to the AML approach. The resulting estimator, known as the conditional expectile, gives a quantile-like extension of the expectation. However, the ALS approach does not estimate the conditional quantiles for counts in a strict sense, so the interpretation is difficult.

Finally, the existence of zero inflation in count data is highlighted in various disciplines (Lambert, 1992; Bohning et al., 1999; Agarwal et al., 2002; Lord et al., 2005); however, to the best of our knowledge, no reliable QR models have been developed for count response with excess zeros. In this dissertation, we propose a novel method for modeling the quantiles of counts. In particular, we focus on the extension to ZI settings that incorporate ZI models in QR.

1.4 Longitudinal Setting

Longitudinal or panel study designs with count responses can also be subject to zero inflation. As noted in Feng and Zhu (2011), ignoring the within-cluster correlation of longitudinal data will lead to loss of efficiency and incorrect inference of the regression coefficients. For the estimation of the conditional mean structure, 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 (2004) framed the approach for finding the MLEs in marginal ZIP regression models by using generalized estimating equations (GEEs). 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 1.1, but where the count distribution must belong to the exponential dispersion family (Jørgensen, 1987). Let Z_{ij} be the indicator variable that Y_{ij} came from the degenerate distribution at 0. Under the assumption of independence, the complete data loglikelihood can be separated and estimation is conducted using an EM algorithm; see Hall and Zhang (2004) for how this procedure is explicitly defined. In the above, the formulas used in estimation have the form of (weighted) GEEs with working correlation matrix equal to the identity matrix. Hall and Zhang (2004) 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, the authors advocate using the GEE-1 approach of Liang et al. (1992), which treats the first and second moment parameters orthogonally. Finally, Iddi and Molenberghs (2013) extended the framework of Hall and Zhang (2004) and presented a marginalized ZI overdispersed model for correlated data.

The basic framework of the conditional model approach extends the traditional ZI models defined in Section 1.1 by including random effects in the estimation of $g(\eta_i)$ and $h(\pi_i)$ for the i^{th} cluster, $i = 1, \dots, M$. The random effects are assumed to be independent normal random variables as in the longitudinal literature. This approach was first considered in Hall (2000) for ZIP and ZIB regression with random intercepts. In particular, the paper considered a random effect for the count regression component. For the ZIP regression with a random effect, let $\mathbf{y}_i \in \mathbb{R}^{n_i}$ be a vector of responses for the i^{th} cluster, $i = 1, \dots, M$. Assume M independent random variables $u_1, \dots, u_M$ are from the standard normal distribution. Conditional on a random effect u_i, the random variable Y_{ij} associated with the observation y_{ij}, $j = 1, \dots, n_i$, follows a ZIP distribution.

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } \pi_{ij}; \\ Poisson(\lambda_{ij}), & \text{with probability } (1 - \pi_{ij}). \end{cases} \tag{1.6}$$

Hence, Hall (2000) obtained the following mixed effect model for the conditional mean, $\boldsymbol{\lambda}_i$

$$log(\boldsymbol{\lambda}_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \sigma u_i \tag{1.7}$$

while the logistic regression model for the mixing probability π is the same as in traditional ZI models. Parameter estimation was conducted by the EM algorithm as in the fixed effects case. However, the computation is more complicated after the inclusion of the random effects. To deal with the challenge, both the indicator variables for the state of the process and the random effects were regarded as missing data. The EM algorithm with Gaussian quadrature was employed to maximize the log likelihood for the ZIP model with random effects.

Wang et al. (2002) extended the above model and included the random effects for both the logistic regression portion and the count regression portion. The model assumed both random effects to be independent normal random variables. Inspired by the generalized linear mixed models (GLMM), they obtained the best linear unbiased estimator (BLUE) via a penalized likelihood function. The residual maximum likelihood (REML) method with an EM algorithm was used for estimation. A similar extension in Fang et al. (2016) developed a hierarchical multilevel ZINB regression model with random effects in both the count regression and zero-inflated portion of the model.

The above conditional model approaches are parametric estimation routines and covariates are assumed to have linear effects on the link function, instead of the observed responses. More recently, the research on semi-parametric estimation explored the problem with more flexibility on the assumption. A smooth function can be included to allow for a nonlinear effect of one continuous covariate. Feng and Zhu (2011) considered a semi-parametric estimation by introducing a nonlinear relationship for one particular covariate in the ZIP setting with longitudinal data. The two-component mixture model becomes

$$\begin{cases} log(\boldsymbol{\lambda}_{ij}) = \mathbf{X}_{ij}^T\boldsymbol{\beta} + u_i + k(T_{ij}) \\ logit(\pi_i) = \mathbf{W}_i^T\boldsymbol{\alpha} \end{cases} \tag{1.8}$$

The smooth function can be estimated by various nonparametric smoothing methods. Feng and Zhu (2011) considered penalized splines with flexible choices of knots and penalty terms. A Monte Carlo EM algorithm for the penalized log-likelihood is used to estimate parameters and smoothing function, $k(T_{ij})$. In general, this semi-parametric estimation routine is more robust but requires more computation resources. The model usually includes only one smoothing function. Computationally, the `glmmTMB` package (Brooks et al., 2017) discussed in later chapters has made estimation of ZI count regression models in the presence of mixed effects quite accessible for a variety of discrete distributions for the count component.

The extension of QR models to longitudinal data (or panel data as it is called in the field of econometrics) is almost exclusive to continuous responses. Inspired by a classical random effects model, Koenker (2004) discussed QR for panel data by including a vector of individual effects in the linear QR setting. To alleviate the increased variability in the estimation, regularization or shrinkage methods drive these individual effects toward a common value. In particular, a class of penalized estimators are proposed where l_1 regularization is introduced to the check loss function.

While theoretical and methodological research in the last 40 years have been addressing important generalizations of the original approach (Koenker, 2017), the literature on the analysis of discrete data remains open to challenges and possibilities. In many applications, practitioners face the limitations of classical parametric models, where the effect of a treatment variable can be heterogeneous throughout the conditional distribution of the count variable, but policy recommendations can only be based on average effects. See Cameron and Trivedi (2013) for a detailed summary of econometric analysis with count data.

An important exception in the literature is the recent work by Chernozhukov, Fernández-Val, Melly, and Wüthrich (2020), who investigate inference for quantile functions, offering simultaneous confidence bands for discrete response variables. While they consider the analysis of cross-sectional data instead of longitudinal data, their work illustrates the increasing importance of flexible methods for count data.

The literature on panel quantiles includes just a few papers, and two main methods for smoothing the discrete objective functions are considered. The original work of Machado and Santos Silva (2005) introduced a jittering approach to smooth the count response variable. Lee and Neocleous (2010) proposed a Bayesian approach, and Chernozhukov, Fernández-Val, and Weidner (2017) develop an approach based on distribution regression. Harding and Lamarche (2019a) extend the jittering approach to longitudinal data without zero inflation and Wang, Wu, Zhao, and Zhou (2020) propose an estimator for time-varying coefficients using a quadratic inference function approach within a quantile framework. The estimator proposed in this paper is different than existing approaches for two important reasons. First, existing quantile regression approaches have not been developed for zero-inflated (ZI) models for longitudinal data. Second, we consider estimation of the conditional mean model in the first step, rather than considering a quantile regression model as in Padellini and Rue (2019a). Therefore, the proposed methodology allows practitioners to estimate a class of models with subject heterogeneity, without considerations on the minimum number of repeated observations per subject as in panel data quantile regression models (Harding and Lamarche, 2019a).

In a longitudinal setting where the response is discrete, there are two main works available. Harding and Lamarche (2019b) proposed a penalized QR model for count data. The novel estimator combines the jittering approach (Machado and Santos Silva, 2005) with the penalized method (Koenker, 2004; Lamarche, 2010), and applied their method to a panel of transactions. Another work by Chernozhukov et al. (2017) incorporated the distribution regression (DR) method first introduced by Williams and Grizzle (1972). This method regressed the empirical cumulative distribution function on the covariates and also modeled the canonical parameters conditional on covariates.

The rest of this dissertation is organized as follows: In Chapter 2, we develop a novel modeling strategy that synthesizes zero-inflated models with quantile regression models. The performance of our method is characterized through extensive Monte Carlo simulations. Finally, we illustrate the method with an application to the Oregon Health Insurance Experiment data. In Chapter 3, we extend the proposed model to the setting of longitudinal data with zero-inflated count responses. A simple three-step approach to estimate the effects of covariates on the quantiles of the response variable is introduced in this chapter. We then present a simulation study to show the small sample performance of the estimator. Finally, we illustrate our model using the RAND Health Insurance Experiment data. In Chapter 4, we explore the Combined Pharmacotherapies and Behavioral Interventions (COMBINE) data with the proposed model from Chapter 3. In particular, we analyze the lasting effects of the therapies, accounting for the socioeconomic and policy factors. Finally, we summarize the models and their applications in Chapter 5.

Chapter 2 Novel Strategies for Quantile Count Regression

2.1 Introduction

Quantiles provide important information in characterizing a distribution, just like the expectation. However, the expectation summarizes the central tendency, while the quantiles can describe the complete distribution. Thus, the utilization of quantiles is preferred if the goal is to explore aspects other than the average. Of all the possible choices, perhaps the most famous quantile is obtained at $\tau = 0.5$, such that the resulting 50th quantile is the median. In fact, for certain distributions, the median equals the expectation. For example, the normal distribution and the t distribution with a degree of freedom greater than one. Hence, the quantile can be considered as an extension of the expectation.

The analysis of the quantiles, especially the median and the quartiles, has a long history (Galton, 1883), but one of the most significant advancements was due to Koenker and Bassett (1978) for developing linear quantile regression (QR). The goal of linear QR is to explore the conditional quantiles of the response, Y, given values of the independent variables, $\mathbf{X}$. Since the seminal work by Koenker and Bassett (1978), countless theoretical results and methodological advancements have been made in this area. Numerous real data analyses have also been informed by QR. For a detailed discussion of QR literature and recent advancements, see Yu et al. (2003), Koenker (2005) and Koenker et al. (2017).

The classic QR model does not assume a particular distribution for the response, and is mainly studied in a nonparametric framework; see Koenker and Bassett (1978), Takeuchi et al. (2006) and Chaudhuri (1991). More recently, parametric QR has been studied from a Bayesian perspective (Yue and Rue, 2011), where one of the most important advancements is the introduction of the asymmetric Laplace distribution (ALD) by Yu and Moyeed (2001). In a Bayesian set-up, the use of the ALD is critical since a working likelihood is required. For a frequentist procedure, the ALD also provides a useful tool for likelihood-based inference, as in Geraci and Bottai (2007).

Computationally, there is usually no closed-form solution to the objective function in

QR modeling, so results are obtained by numerical algorithms. Koenker and Bassett (1978) showed that the estimation can be transformed into a linear programming (LP) problem, where multiple algorithms are available. To date, the most popular choice is the simplex algorithm (Barrodale and Roberts, 1978). When the sample size is moderate, the Simplex algorithm is efficient and fast. The interior point algorithms are preferred for large sample sizes (Portnoy and Koenker, 1997) or nonlinear modeling (Koenker and Park, 1996). For more recent advancements in computational resources, see Geraci (2014, 2016).

In practice, researchers often need to model data where the response is counts of interest; for example, the counts of failures during a manufacturing process (Lambert, 1992), the number of motor vehicle crashes (Lord et al., 2005), or the number of housing units in a census block (Young et al., 2017). QR was originally developed for continuous response, and extending to discrete response is non-trivial. This issue arises from the fact that the discrete response gives a nondifferentiable objective function. Consequently, the traditional routine for estimation does not guarantee convergence. To model the conditional quantiles of discrete response, a pragmatic framework is to approximate discrete data with continuous distributions that share certain characterizations.

A feature of count data that is commonly encountered is excess zeros relative to an assumed count model. Count data with excess zeros are commonly found in many areas, including industry (Lambert, 1992), epidemiology (Bohning et al., 1999), ecology (Agarwal et al., 2002), transportation (Lord et al., 2005) and insurance (Baetschmann and Winkelmann, 2012). This phenomenon is formally known as zero-inflation. When the data contains a higher proportion of zero counts than expected under a model, this leads to unstable estimation and misleading inference. To account for excess zeros, zero-inflated (ZI) models analyze the data as a mixture of two components: one component is the goal process characterized by the non-ZI count model while the other component is a degenerate distribution at zero. The seminal work of Lambert (1992) introduced the zero-inflated Poisson (ZIP) regression model. The model was characterized by the first component considered as an imperfect state, where random zero counts occur by the Poisson distribution and the second component considered as a perfect state, where a structural zero is the only possibility.

Since the expectation and the variance of a Poisson distribution are equal, the ZIP model is most efficient when the variability matches the expected value. The zero-inflated negative binomial (ZINB) regression model is popular when the data are overdispersed other than from excess zeros. The negative binomial distribution has separate mean and dispersion parameters, hence, the ZINB is more flexible in situations when the variability is greater than expected. Following the modeling strategies in Lambert (1992), Greene (1994) considered modification of the negative binomial distribution to accomodate zero-inflation, and discussed the distinction between zero-inflation and overdispersion. Ridout et al. (2001) discussed ZINB models with inferences based on a score test. For more recent works on ZINB models, see Yau et al. (2003) and Mwalili et al. (2008).

The estimation routine for ZI modeling is commonly conducted via the expectation-maximization (EM) algorithm (Dempster et al., 1977) as in the original work by Lambert (1992). Another alternative is Newton-Raphson algorithm, which is faster to convergence than the EM algorithm; however, Newton-Raphson algorithm could fail to converge, as noted by Lambert (1992).

Inferential aspects about ZI models have also been studied in the literature. The most fundamental question is whether ZI count regression shows improvements over the corresponding count regression without ZI adjustments. This is equivalent to testing the presence of zero-inflation. To conduct such tests, researchers often utilize score tests (van den Broek, 1995; Jansakul and Hinde, 2002, 2008), a boundary likelihood ratio test (Hilbe, 2011), or Vuong's non-nested test based on likelihood ratio (Vuong, 1989).

Diagnostics based on residual analysis also provide straightforward, yet insightful information in ZI count regression. Traditionally, researchers assess Pearson or deviance residuals given by the models. More recently, the randomized quantile residuals proposed by Dunn and Smyth (1996) are used for assessing model fitting in ZI count regression. For example, Young et al. (2017) utilized randomized quantile residuals in developing census frames for the 2020 Census.

The popularity of ZI count regression models is highlighted by their application in various disciplines (Lambert, 1992; Bohning et al., 1999; Agarwal et al., 2002; Lord et al., 2005; Young et al., 2017), however, most ZI models focus on the mean structure just like the linear regression analyses. To the best of our knowledge, no reliable QR models have been developed for count response with excess zeros. In this paper, we proposed a three-step method for modeling the quantiles of counts. In particular, we focus on the extension to ZI settings that incorporate ZI models in QR.

Our work is focused on an application involving the Oregon health insurance data from Finkelstein et al. (2012). This randomized control trial explored the influence of multiple covariates on the health care utilization in Oregon. The response of interest is the number of visits to doctor so it is a discrete variable. An excess number of zeros is also present in the data.

The rest of the paper is organized as follows. Section 2.2 discusses quantile regression for count data, where the main method is by Machado and Santos Silva (2005). Section 2.3 introduces an alternative approach for quantile count regression proposed by Padellini and Rue (2019b). In this section, the new approach from a frequentist perspective is discussed in detail. Section 2.4 concerns the extension of our proposed model to ZI setting. Section 2.5 reports simulation results for different methods. Section 2.6 provides the empirical application to the Oregon health insurance data. Section 2.7 gives conclusions and discussion for future research. Section 2.8 provides mathematical details for derivations, additional figures and further details for simulation results in the Appendix.

2.2 Quantiles for Counts

Let Y be a random variable with cumulative distribution function (CDF) $F_Y(y) = P(Y \leq y)$. The τ^{th} quantile is defined as:

$$Q_Y(\tau) = inf \left\{ y : F_Y(y) \geq \tau \right\}, \tag{2.1}$$

where $0 < \tau < 1$ is the quantile level. When Y is a continuous random variable with strictly-increasing CDF F_Y, then,

$$Q_Y(\tau) = F_Y^{-1}, \tag{2.2}$$

where $Q_Y(\tau)$ is a strictly-increasing function of τ. Hence,

$$F_Y[Q_Y(\tau)] = P[Y \leq Q_Y(\tau)] = \tau \tag{2.3}$$

As in the case of classic linear regression, the goal of linear QR is to explore the conditional quantiles of the response, Y, given values of the independent variables, $\mathbf{X}$. Specifically,

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^t\beta\left(\tau\right), \tag{2.4}$$

where the regression coefficients, $\beta\left(\tau\right)$ depends on the quantile level, τ. However, when Y is a discrete random variable, there is no one-to-one relationship between $Q_Y(\tau)$ and τ since both CDF and quantile functions are step functions. Furthermore, the non-differentiability inhibits the extension of the optimization routine as in the continuous case. Lastly, the linearity assumption does not hold for most problems when the response is a count.

A popular approach proposed by Machado and Santos Silva (2005) is to construct a continuous variable by jittering on the original counts. For the computational routine to work, an independent random variable that follows a continuous uniform distribution in $[0, 1)$, is generated. The random noise is then added to the original count response to transform it into a continuous variable. The traditional QR methods can then be applied to the updated data.

The main advantage of the jittering-based approach is that all of the existing QR methods can be applied easily after the transformation. However, this method does not guarantee the conditional quantiles for the new response to be the same as the conditional quantiles for the original response. Another concern with the jittering-based approach is quantiles crossing. By definition, the conditional quantile function $Q_Y(\tau|\mathbf{x})$ is a nondecreasing function of τ for any $\mathbf{x}$. This implies that, for $\tau_1 > \tau_2$,

$$Q_Y(\tau_1|\mathbf{x}) > Q_Y(\tau_2|\mathbf{x}) \tag{2.5}$$

Hence, quantiles crossing should be considered as an indicator of inaccurate estimation. In practice, however, the jittering-based procedure tends to incur quantiles crossing. This problem originates from the estimation routine in the traditional QR; furthermore, the random terms added to the count worsens the situation.

An alternative approach for QR with count data is the asymmetric maximum likelihood (AML) estimator proposed by Efron (1992). This estimator arises from the optimization of a smoothed objective function as given in Koenker and Bassett (1978) and hence is straightforward to interpret. However, the computation routine requires the quantile level, τ, to be greater than the proportion of zeros in the data. Thus, their method does not work for ZI settings. Newey and Powell (1987) proposed an asymmetric least squares (ALS) estimation that is akin to the AML approach. The resulting estimator, known as the conditional expectile, gives a quantile-like extension of the expectation. However, the ALS approach does not estimate the conditional quantiles for counts in a strict sense, so the interpretation is difficult.

In order to overcome these issues and obtain reliable estimation, Padellini and Rue (2019b) introduced another approximation routine based on mathematical interpolation. This is the basis for our approach discussed in the next section.

2.3 Three-Step Quantile Count Regression

Interpolation

To approximate the quantiles for discrete response, another strategy is by interpolation. In numerical analysis, interpolation is a method to construct new data points within the range of a finite set of data points. This links the discrete variable to a continuous distribution, where the CDFs match at integer values. Thus, the discrete response can be viewed as generated by a continuous counterpart of the original discrete distribution.

To achieve this goal, the CDF of a discrete variable is required to satisfy the following condition:

$$F_Y\left(y;\theta\right) = P\left(Y \leq y\right) = k\left(\lfloor y \rfloor;\theta\right) \tag{2.6}$$

where k is a continuous function and $\lfloor \quad \rfloor$ is the floor operator so $\lfloor y \rfloor$ is the greatest integer that is less than or equal to y. When y is an integer, $\lfloor y \rfloor = y$.

To obtain the corresponding continuous distribution, we remove the floor operator. The CDF of a discrete random variable, Y, then becomes the CDF of a continuous counterpart, $\tilde{Y}$ since:

$$F_Y\left(y;\theta\right) = k\left(\lfloor y \rfloor;\theta\right) \Rightarrow k\left(y;\theta\right) = F_{\tilde{Y}}\left(y;\theta\right), \tag{2.7}$$

Ilienko (2013a) showed that the most common distributions for modeling counts, including binomial, Poisson and negative binomial distribution, satisfy this condition.

Proposition 2.3.1. *Suppose $Y \sim$ Poisson (λ) and $\tilde{Y}$ is the continuous counterpart of Y by interpolation, then $\tilde{Y}$ follows a continuous Poisson distribution with CDF,*

$$F_{\tilde{Y}}(y) = \frac{\Gamma(y+1,\lambda)}{\Gamma(y+1)}, y \geq 0, \tag{2.8}$$

where $\Gamma(y+1,\lambda)$ is the upper incomplete gamma function.

This relationship is illustrated by the following graph for several values of λ. The step functions are the theoretical CDFs of the discrete random variable, $F_Y(y)$. The curves are $\dfrac{\Gamma(y+1,\lambda)}{\Gamma(y+1)} = F_{\tilde{Y}}(y)$, the CDFs of the continuous random variables. Notice that the two functions match at the non-negative integer values.

Proposition 2.3.2. *Suppose Y is the number of successes to get r failures in a series of independent Bernoulli trials with a success probability, p. By definition, $Y \sim$ Negative Binomial (r,p). Let $\tilde{Y}$ be the continuous counterpart of Y by interpolation, then $\tilde{Y}$ follows a continuous negative binomial distribution with CDF,*

$$F_{\tilde{Y}}(y) = I_{1-p}(r, y+1), \tag{2.9}$$

where $I_x(a,b)$ is the regularized incomplete Beta function.

A similar illustration of this relationship is provided in the Appendix 2.8. Proofs of proposition (2.3.1) and (2.3.2) are also provided in the Appendix 2.8.

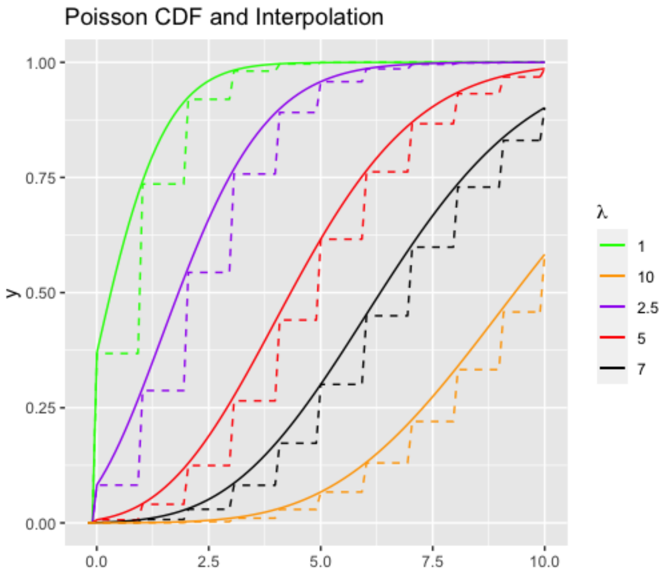


Figure 2.1: Poisson CDF(dashed line) and continuous Poisson CDF(solid curve).

Three-Step Approach

Based on the continuous counterpart of the generating distribution, Padellini and Rue (2019b) proposed a model-aware approach for QR with discrete response in Bayesian setting. Inspired by their method, we propose a three-step quantile count regression (TQCR) in frequentist setting.

Given data $(Y_i, \mathbf{X}_i)$, suppose that $Y_i|\mathbf{X}_i \sim F(y_i;\theta_i)$, where $\theta_i = E(Y_i|\mathbf{X}_i)$, the conditional mean.

1. **Step 1**: The conditional mean structure is modeled as

$$\theta_i = g_1(\mathbf{X}_i), \tag{2.10}$$

where g_1 is an invertible function determined by the problem. This is analogous to fitting a regression function in a linear mixed model, or to specifying a link function in a generalized linear (mixed) model.

2. **Step 2**: The conditional quantile $Q_Y(\tau|\mathbf{X}_i)$ is mapped to the conditional mean parameter, θ_i as

$$Q_Y(\tau|\mathbf{X}_i) = g_2(\theta_i), \tag{2.11}$$

where g_2 is given by the relation between the mean and quantile in the specified distribution.

3. **Step 3**: The resulting composite $g_2 \circ g_1 = g_2(g_1)$ maps the conditional quantile of Y to independent variables, $\mathbf{X}$. That is,

$$Q_Y(\tau|\mathbf{X}_i) = g_2 \circ g_1(\mathbf{X}_i), \tag{2.12}$$

The composite $g_2 \circ g_1 = g_2(g_1)$ is usually non linear. As a result, non linear least square (NLS) method is employed for model fitting in this step. It can be seen from the above that, the three-step approach can make full use of existing methods to estimate the mean function in Step 1. This applies to a continuous response variable as well as a discrete response variable. Techniques for fitting the mean structure include parametric, semi-parametric and non-parametric techniques. Thus, the researcher is afforded considerable flexibility to choose the most appropriate model for their problems. Examples of different fitting methods for Step 1 are given in the Appendix 2.8.

Suppose the conditional distribution of the response variable, $Y|\mathbf{X} = \mathbf{x}$ follows a discrete distribution that satisfies condition (2.6). Assume further that the discrete distribution is fully parametric with parameter vector, θ. In practice, the modeling strategy is implemented via the following steps described in Procedure 1.

Procedure 1 Modeling Procedure for Quantile Regression with Count Data

(1) Select a discrete distribution and estimate the parameters θ by fitting a corresponding GLM to the data. Since the most common choices for discrete data are the Poisson distribution and negative binomial distribution, this step can be conducted by fitting their respective regression model.

(2) Plug in the estimated parameters $\hat{\theta}$ to the interpolated CDF of the specified distribution. This estimates the CDF of the corresponding continuous distribution, $\hat{F}_{\tilde{Y}}(y)$.

(3) Obtain the conditional quantiles of the continuous counterpart, $\hat{Q}_{\tilde{Y}}(\tau|\mathbf{X} = \mathbf{x})$. This is done by numerically solving for the value,

$$\hat{Q}_{\tilde{Y}}(\tau|\mathbf{X} = \mathbf{x}) = argmin \left\{ y : \hat{F}_{\tilde{Y}}(y) \geq \tau|\mathbf{X} = \mathbf{x} \right\} \tag{2.13}$$

The unique value is guaranteed by the identifiability for the continuous distribution.

(4) Round the estimated quantiles up to the next integer, yielding the conditional quantiles for the original discrete distribution.

(5) Fit a non linear function to the estimated quantiles of the continuous counterpart and the predictor variables, $\mathbf{X}$. This models the composite $g_2 \circ g_1$ and estimates the quantile regression coefficients, $\boldsymbol{\beta}(\tau)$.

Hence, the above provides a new approach for fitting a QR model with discrete responses and for estimating the quantile regression coefficients, $\boldsymbol{\beta}(\tau)$.

Distribution Regression

Another option to model the composite $g_2 \circ g_1$ incorporates the distribution regression (DR) method first introduced by Williams and Grizzle (1972) for the ordered response.

Chernozhukov et al. (2018) considered DR estimator in a Poisson regression setting, given by

$$\hat{F}_{Y|\mathbf{x}}(y|\mathbf{x}) = \Lambda_y(\mathbf{x}'\hat{\boldsymbol{\beta}}(y)), \tag{2.14}$$

where

$$\hat{\beta}(y) = \underset{\beta \in \Theta}{\arg\max} \sum_i I_{\{Y_i \leq y\}} \cdot ln\left[\Lambda_y(\mathbf{X}_i'\beta)\right] + I_{\{Y_i > y\}} \cdot ln\left[1 - \Lambda_y(\mathbf{X}_i'\beta)\right], \tag{2.15}$$

When $y = Q_Y(\tau)$ in (2.15), the DR coefficient $\hat{\beta}$ depends on the quantile level τ, and hence $\hat{\beta}(\tau)$ can be interpreted as the regression coefficients as in a QR setting.

As can be seen in the above objective function (2.15), the DR method estimates the CDF as well as the canonical parameter when modeling the response conditional on covariates. This is similar to the idea of the three-step approach. Hence, DR can be combined naturally into a three-step QR framework.

The above routines can be used in a QR for count response. In this situation, one advantage of our approach for discrete response variable is that, compared with Machado and Santos Silva (2005), our method is less affected by quantiles crossing. A graphical comparison is provided by plots 2.10 in the Appendix 2.8. Another advantage of the three-step approach is that it directly reflects characteristics of the data. This is particularly effective when, for example, data show evidence for a specific distribution or possess excess zeros relative to such a distribution.

2.4 Three-Step Quantile Count Regression for Zero-inflated Data

When the response is discrete, we cannot employ traditional QR methods, and even the jittering-based approach has its own limitations. If the dataset also contains a certain proportion of zero counts, the estimation of the conditional quantiles becomes more challenging. In this section, we extend the TQCR to the setting where the data possess excess zeros. Following the terminology in Lambert (1992), we emphasize the distinction in estimating the conditional quantiles for the Complete Process (the complete data is generated by one data generating mechanism) versus that for the Count Process (the non-ZI count distribution). Estimation of the Zero Process itself is not intriguing, since the Zero Process is a degenerate distribution that only generates zero. In that case, there is no need to explore the complete distribution.

Zero-Inflated Model

ZI models can be considered as a special case of a mixture model: one component is a point mass at zero and the other component is a discrete distribution. The first component, considered a perfect state where the events of interests cannot occur, is the source of excess zeros in the data; the second component, considered an imperfect state where events occur according the assumed distribution, is often the focus of interest to researchers.

The probability of being in the perfect state, or the probability of a structural zero, p_0, is modeled by a logistic regression:

$$logit(p_0) = log\left(\frac{p_0}{1-p_0}\right) = \mathbf{z}^T\mathbf{w}, \tag{2.16}$$

where $\mathbf{z}$ is the vector of covariates. The conditional mean of the count process, $E[Y|\mathbf{X} = \mathbf{x}] = \theta$, is modeled by a discrete distribution via a known link function, $g(\theta) = \mathbf{x}^T\beta$. In practice, a log link function is usually assumed, thus,

$$log(\theta) = \mathbf{x}^T\beta, \tag{2.17}$$

The most common choices for the discrete distribution in the ZI model are the Poisson distribution and the negative binomial distribution. Note that the covariates for the count process can be the same set of variables as the covariates for modeling the proportion of being in the perfect state. In that case, $\mathbf{z} = \mathbf{x}$ in equations (2.16) and (2.17).

In order to establish the continuous counterpart of a ZI count distribution, we first note the following relationship between the CDF of a discrete random variable and its ZI version:

$$F_{\tilde{Y}}(y) = \begin{cases} p_0 + (1-p_0) \cdot F_Y(y), \text{ if } y \geq 0, \\ 0, \text{ if } y < 0. \end{cases} \tag{2.18}$$

We then have the following proposition, the proof of which is in Appendix 2.8.

Proposition 2.4.1. *$F_Y(y)$ is a valid cumulative distribution function.*

We now derive the continuous counterparts of the ZIP and ZINB distributions.

First, suppose $Y \sim Poisson(\lambda)$ and $\tilde{Y}$ is the ZI counterpart of Y. That is, $\tilde{Y} \sim ZIP(\lambda, p_0)$.

$$F_{\tilde{Y}}(y) = p_0 + (1-p_0) \cdot F_Y(y) = p_0 + (1-p_0) \cdot \frac{\Gamma(\lfloor y \rfloor + 1, \lambda)}{\Gamma(\lfloor y \rfloor + 1)}, y \geq 0. \tag{2.19}$$

The continuous counterpart is obtained by removing the floor function $\lfloor \cdot \rfloor$, as in the previous examples. Thus,

$$F_{\tilde{Y}'(y)} = p_0 + (1-p_0) \cdot \frac{\Gamma(y+1, \lambda)}{\Gamma(y+1)}, y \geq 0. \tag{2.20}$$

As one can see, this extension only utilizes the property of CDF.

This extension to ZI model also works for negative binomial(r,p) with proportion of zero-inflation equals p_0. Suppose that $Y \sim$ negative binomial (r, p) and $\tilde{Y}$ is the Zero-Inflated counterpart of Y, then, $\tilde{Y} \sim ZINB(r, p, p_0)$, and we have the CDF,

$$F_{\tilde{Y}}(y) = \begin{cases} p_0 + (1-p_0) \cdot I_{1-p}\left(r, \lfloor y \rfloor + 1\right), y \geq 0, \\ 0, \text{ if } y < 0. \end{cases} \tag{2.21}$$

Again, the continuous counterpart is obtained by removing the floor function $\lfloor \cdot \rfloor$,

$$F_{\tilde{Y}'}(y) = p_0 + (1-p_0) \cdot I_{1-p}\left(r, y+1\right) = \frac{B(r, y+1, 1-p)}{B(r, y+1))}, y \geq 0. \tag{2.22}$$

A similar illustration of this relationship is provided in Figure 2.2.

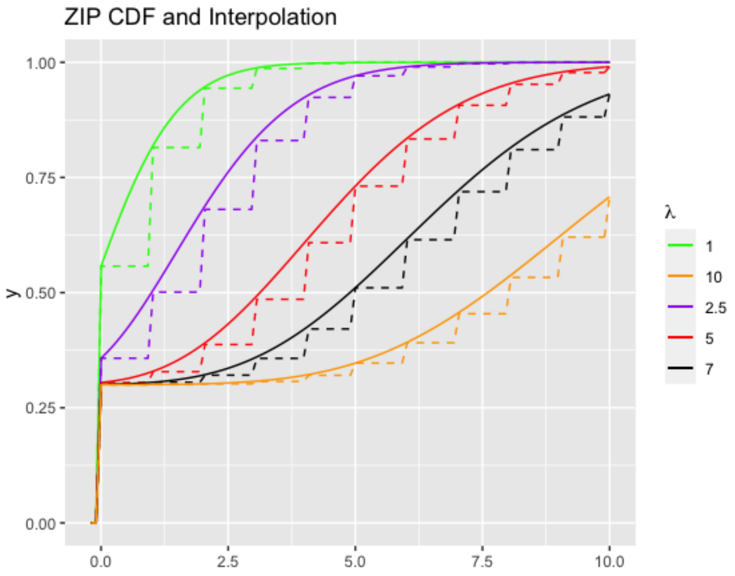


Figure 2.2: ZIP CDF (dashed line) and interpolation (solid curve).

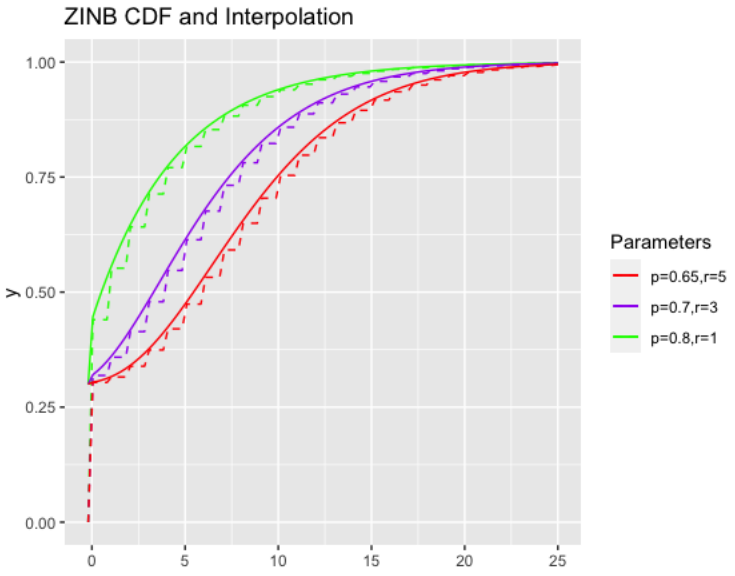


Figure 2.3: ZINB CDF (dashed line) and interpolation (solid curve).

Identifiability

To guarantee unique solutions in estimating model parameters, it is important to construct a model that is identifiable. Li (2012) showed that ZIP models are identifiable. Based on Li (2012), we stated the following proposition for the continuous counterpart of ZIP models.

The proof is included in 2.8.

Proposition 2.4.2. *For the continuous counterpart of a ZIP model with parameter λ and proportion of zero-inflation p,*

$$f(y; p_1(x), \lambda_1(x), x) = f(y; p_2(x), \lambda_2(x), x) \textit{ implies } \lambda_1(x) = \lambda_2(x) \textit{ and } p_1(x) = p_2(x).$$

Hence, the continuous counterpart of ZIP model is identifiable. Identifiability for the ZINB model is still an open problem. Thus, we do not establish identifiability of the continuous counterpart of the ZINB model in the present work.

Zero-Inflated Quantile Count Regression: Complete Process

We now extend our approach to model the quantiles of ZI count regression data. In the first scenario, we define the quantiles of interest as the quantiles of the Complete Process. Hence, the analysis treats the entire dataset as a single entity.

Given data (Y_i, $\mathbf{X}_i$), suppose that $Y_i|\mathbf{X}_i \sim F(y_i; \theta_i, \pi_i)$, where $\theta_i = E(Y_i|\mathbf{X}_i)$ is the conditional mean for the Count Process, and π_i is the probability that the response is from the Zero Process. In this case, the extension to the ZI dataset is natural. Since the three-step approach is based on the interpolation of the CDF, it only requires replacing the CDF of the discrete distribution with that of the ZI version in equation (2.18). This introduces a modeling strategy for the empirical distribution.

Zero-Inflated Quantile Count Regression: Count Process

In some situations, researchers are only concerned with the underlying count distribution in the dataset without effects from excess zeros. In this case, the ZI model assumes the data arises as a mixture of a discrete distribution and a degenerate distribution at 0. Structural zeros are considered contamination while the quantiles of interest are for the Count Process. Hence, the first goal is to distinguish these two mixture components.

The steps of the modeling strategy for a ZI data is described in Procedure 2.

Procedure 2 Modeling Procedure for Quantile Count Regression with Zero-Inflated Data

(1) Fit a ZI count regression model to the data. Obtain estimates for the parameters $\boldsymbol{\theta}$ of the Count Process and the probability of ZI, π_i.

(2) Plug in the estimated parameters $\hat{\boldsymbol{\theta}}$ to the interpolated CDF of the discrete distribution for the Count Process. This estimates the CDF of the corresponding continuous distribution for the Count Process, $\hat{F}_{\tilde{Y}}(y)$.

(3) Obtain the conditional quantiles of the continuous counterpart, $\hat{Q}_{\tilde{Y}}(\tau|\mathbf{X})$. The unique value is guaranteed by the (assumed) identifiability for the continuous distribution.

(4) Round the estimated quantiles up to the next integer, yielding the conditional quantiles of interest.

(5) Fit a non linear function to the estimated quantiles of the continuous counterpart and the predictor variables, $\mathbf{X}$. This models the composite $g_2 \circ g_1$ and estimates the quantile regression coefficients $\boldsymbol{\beta}(\tau)$ for the Count Process.

From the above description, it can be seen that the interpolation is on the CDF of the discrete distribution for the Count Process instead of the entire data. When the goal is to explore the Count Process, this fits a QR model for the quantiles of interest.

In the presence of zero-inflation, our approach shows much better results compared with the jittering-based approach. This is due to the fact that jittering-based approach does not distinguish the Count process from Zero process. The three-step approach, on the other hand, can tackle this problem. This is accomplished in the first step, where researchers can choose existing methods to analyze the ZI model. This will be the focus of the following sections.

Distribution Regression for Zero-Inflated setting

Inspired by the DR estimator in Chernozhukov et al. (2018), we modified the objective function in equation (2.15) for a ZI setting. When the estimation is on the Complete Process, the modification is the same as the extension from a discrete distribution to the ZI model. Hence, the modeling requires solving equation (2.15) with the CDF of the ZI counterpart, (2.18).

Another situation requires the distinction between two generating mechanisms. In this situation, the objective function for a ZI setting is given by:

$$\begin{aligned}\sum_i I_{\{Y_i\leq y\}} &\left[ln\left(\Lambda_y(\mathbf{X}_i'\boldsymbol{\beta})\right) I_{\{Y_i>0\}} + \left(p_i + (1-p_i)\cdot ln\left(\Lambda_y(\mathbf{X}_i'\boldsymbol{\beta})\right)\right) I_{\{Y_i=0\}}\right] \\ &+ I_{\{Y_i>y\}} ln\left[1-\Lambda_y(\mathbf{X}_i'\boldsymbol{\beta})\right],\end{aligned} \tag{2.23}$$

where p_i is the probability that an observation comes from the zero process. It can be seen that only the first component in objective function (2.15) is adjusted for a ZI setting.

If a positive count is observed, then the observation must come from the count process and hence $p_i = 0$. The first component in the objective function (2.23) becomes,

$$I_{\{Y_i\leq y\}}\cdot\left[ln\left(\Lambda_y(\mathbf{X}_i'\boldsymbol{\beta})\right) + (0 + 1\cdot ln\left(\Lambda_y(\mathbf{X}_i'\boldsymbol{\beta})\right))\cdot 0\right] = I_{\{Y_i\leq y\}}\cdot ln\left[\Lambda_y(\mathbf{X}_i'\boldsymbol{\beta})\right], \tag{2.24}$$

If a zero count from the count process is observed, then $p_i = 0$. The first component in objective function (2.23) becomes,

$$I_{\{Y_i\leq y\}}\cdot\left[0 + (0 + ln\left(\Lambda_y(\mathbf{X}_i'\boldsymbol{\beta})\right))\right] = I_{\{Y_i\leq y\}}\cdot ln\left[\Lambda_y(\mathbf{X}_i'\boldsymbol{\beta})\right], \tag{2.25}$$

In both situations, the first component in (2.23) is the same as the first component in (2.15), that is, the regular situation without zero-inflation.

When zero-inflation exists and a structural zero from the zero process is observed, then $p_i > 0$. The first component in (2.23) becomes,

$$I_{\{Y_i \le y\}} \cdot \left[0 + (p_i + (1-p_i)\cdot ln\left(\Lambda_y(\mathbf{X}_i^{'}\boldsymbol{\beta})\right))\right] = I_{\{Y_i \le y\}} \cdot \left(p_i + (1-p_i)\cdot ln\left[\Lambda_y(\mathbf{X}_i^{'}\boldsymbol{\beta})\right]\right), \tag{2.26}$$

This is in accordance with the CDF in expression (2.18).

Inference

For classic ZI models, we can rely on established asymptotic theory for estimating standard error (SE) and conducting inferences, that is, constructing confidence intervals and testing hypotheses. QR models, on the other hand, employ both asymptotic theory and bootstrapping to obtain the SE.

We proceed to compare different bootstrap methods and their performance in inference. One natural choice is a parametric bootstrap routine, given the fact that the three-step approach estimates the canonical parameters, $\boldsymbol{\theta}$. Nonparametric bootstrap routines, such as pairwise bootstrap and multiplier bootstrap, are easy to implement and more robust to the distributional assumptions.

The first nonparametric bootstrap method considered is the pairwise bootstrap. This method applies the idea of sampling with replacement to the pairs of response variable and independent variables. That is, we obtain a bootstrap sample by sampling with replacement from the pairs $(y_i, \mathbf{x}_i)$ where y_i is the observed response and $\mathbf{x}_i$ is the vector of independent variables associated with observation i, $i = 1, \cdots, n$. An equivalent way to think about the pairwise bootstrap is to re-sample with replacement from the sequence of observation numbers: $1, \cdots, n$.

The other nonparametrc bootstrap method to be considered is the multiplier bootstrap, also known as the weighted bootstrap (Ma and Kosorok, 2005). This method fixes the response and independent variables at the original values, hence the data itself is not re-sampled. Instead, another independent vector of weights, $\mathbf{W} = (W_1, \cdots, W_n)^t$, is generated within each bootstrap sample. Further, this vector of weights must be an i.i.d sample such that $E(W_i) = 1$ and $Var(W_i) < \infty$. Finally, each vector of weights is applied to the estimation routine within each bootstrap run. In summary, the multiplier bootstrap is based on a weighted estimation with different weights sampled under certain conditions.

Compared with asymptotic theory, inference based on bootstrapping is easy to compute and straightforward to interpret. Bootstrap confidence interval (CI) can also be obtained in the same computation. This allows one to further explore the inferential aspects of the three-step approach. Results regarding the performance of different inferential procedures are reported in Section 2.5.

Goodness-of-Fit Assessment and Model Selection

Traditional residuals are commonly based on the discrepancy between the observed values, $\mathbf{y}$ and the fitted values, $\hat{\mathbf{y}}$. In a QR setting, however, traditional residuals cannot be applied directly since the observed values are at different quantile levels. In order to calculate the traditional residuals, it is required to know the true quantiles of the response beforehand. Furthermore, when the response is discrete, traditional residual plots are less helpful since the discreteness usually induces near-parallel curves corresponding to different integer-valued response.

To check the goodness-of-fit(GOF) for QR with discrete response, we utilize the randomized quantile residuals proposed by Dunn and Smyth (1996). When the CDF $F(y)$ is continuous, $F(y_i)$ are uniformly distributed on the unit interval. In this case, the randomized quantile residuals are given by:

$$r_{q,i} = \Phi^{-1}\left\{F(y_i;\hat{\theta})\right\}, \tag{2.27}$$

where Φ^{-1} is the CDF of standard normal distribution. This implies that if the model gives consistent estimates of the parameter $\hat{\theta}$, the distribution of $r_{q,i}$ converges to standard normal. Hence, the normal Q-Q plot can be employed for an illustration of model checking.

The three-step approach simultaneously estimates the parameter θ and the quantile $Q_Y(\tau)$ conditional on the values of covariates. As a result, when the method works in a QR setting, it also consistently estimates the parameter. This equivalence allows assessment of the GOF for the fitted QR model based on the assessment for the parameter. A graphical illustration based on simulated data can be found in Figure 2.12 and 2.13 in the Appendix 2.8.

Because the three-step approach uses interpolation to obtain a continuous counterpart of the discrete response, the resulting CDF is continuous. Thus, the application of randomized quantile residuals is straightforward. It is worth noting that there is a more general definition of randomized quantile residuals if the CDF is not continuous (Dunn and Smyth, 1996).

Multiple inferential procedures can be used to determine the distribution that bests fits the data. Likelihood ratio (LR) test, score test and Vuong' non-nested test can be used for comparison of distributions (van den Broek, 1995; Jansakul and Hinde, 2002; Hilbe, 2011). When LR test is used to test the presence of ZI, a boundary-correction gives more power, see Hilbe (2011). Information criteria, such as AIC and BIC, can also be used for model selection (Hilbe, 2011). In summary, GOF assessment and model selection should be conducted to find the distribution or the structure that best fits the data. As the simulation results indicate in the next section, this step is advantageous to achieve the superb results.

2.5 Simulation Studies

This section reports the results of exclusive Monte Carlo simulations to investigate the performance of our method compared with existing method. The overall performance is measured by the empirical mean integrated squared errors(MISE) of each method. The standard errors of estimators are reported to compare efficiencies.

We adopted similar simulation conditions as in Machado and Santos Silva (2005). In the first two simulation settings, the responses $Y_1, Y_2, \cdots, Y_n$ were generated from Poisson distributions; in the third setting, the responses were generated from negative binomial distribution. Under each simulation setting, data were generated with a proportion of zero-inflation equals $\{0, 0.10, 0.45\}$, respectively. That is, data with no zero-inflation, slight zero-inflation, moderate zero-inflation. Notice that while the proportion of zero-inflation varies, the structure of the discrete distribution stays the same; hence, the stability of the estimators (the values of the estimators do not significantly vary across different zero-inflation settings) indicates good performance with respect to the Count Process.

In the first simulation setting, the conditional mean of the count process is $\mu_i = exp(b_0 + b_1 \cdot x_{1i})$, where the x_{1i}'s were obtained as random draws from a uniform distribution over $(0, 5)$. In the second simulation setting, the conditional mean of the count process is $\mu_i = exp(b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i})$, where the x_{1i}'s and x_{2i}'s were obtained as random draws from a uniform distribution over $(0, 5)$. In the third simulation setting, the conditional mean μ_i of the count process is the same as in the first situation and the variance equals $\mu_i + 0.5 \cdot \mu_i^2$. All experiments were performed with $(b_0, b_1) = (0.7, 0.5)$ for one covariate and $(b_0, b_1, b_2) = (0.7, 0.5, -0.35)$ for two covariates. $N = 5000$ simulations for each case were performed. Within each simulation, samples with size $n \in \{250, 500\}$ were generated. A different set of values for covariates were drawn in each of the 5000 simulations.

We then compare the MISE of different QR models. The MISE is defined as:

$$MISE = \frac{1}{n} \sum_{i=1}^{n} [\hat{g}_\tau(\mathbf{X}_i) - g_\tau(\mathbf{X}_i)]^2, \tag{2.28}$$

This is an SSE-based criterion, so a smaller value of MISE indicates a better fit to the data. Note that in the literature, some researchers report the square root of the empirical MISE, known as empirical root mean integrated squared errors (RMISE).

Overall, when there is no excess zero counts, the three-step approach with NLS routine shows better results compared with the jittering method; when zero-inflation exists, both three-step approaches based on NLS routine and DR routine show better performance than the jittering method.In particular, the values given by the NLS routine only change slightly across different zero-inflation settings. As we pointed out, this indicates that the estimators are more robust to the contamination of excess zeros. When the primary goal is to explore the Count Process, our approach shows great improvement over the existing methods.

Another observation about the DR routine is that this method tends to perform the best when the primary interest is in the Complete Process and the sample has a medium to large size. The method is originally derived from the empirical distribution of the whole dataset, hence the name ”Distribution”. While its performance in a ZI setting is compromised, it provides a good model in exploring the overall structure, which could be beneficial if considering marginalized ZI count regression models. More simulations details regarding the point is provided in the Appendix.

Table 2.1: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as Poisson distribution. Results based on a sample size $n = 250$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-0.279(0.075)	0.642(0.020)	0.206
	TS-DR	0.145(0.366)	0.502(0.106)	2.874
	JB	-0.243(0.162)	0.644(0.042)	0.496
0.25	TS-NLS	0.197(0.071)	0.573(0.019)	0.182
	TS-DR	0.598(0.154)	0.485(0.040)	1.085
	JB	0.240(0.103)	0.569(0.028)	0.313
0.50	TS-NLS	0.645(0.062)	0.510(0.017)	0.176
	TS-DR	0.770(0.074)	0.490(0.021)	0.534
	JB	0.634(0.079)	0.514(0.022)	0.231
0.75	TS-NLS	1.016(0.056)	0.461(0.016)	0.199
	TS-DR	0.962(0.165)	0.482(0.042)	0.813
	JB	0.971(0.073)	0.468(0.020)	0.325
0.90	TS-NLS	1.296(0.052)	0.425(0.014)	0.227
	TS-DR	1.406(0.121)	0.406(0.033)	1.148
	JB	1.246(0.079)	0.432(0.022)	0.530

Table 2.2: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as Poisson distribution. Results based on a sample size $n = 500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-0.281(0.053)	0.642(0.014)	0.172
	TS-DR	-0.023(0.291)	0.550(0.081)	1.697
	JB	-0.238(0.113)	0.644(0.029)	0.385
0.25	TS-NLS	0.197(0.050)	0.573(0.013)	0.142
	TS-DR	0.551(0.160)	0.497(0.040)	0.875
	JB	0.242(0.074)	0.569(0.020)	0.236
0.50	TS-NLS	0.646(0.044)	0.510(0.012)	0.129
	TS-DR	0.780(0.056)	0.487(0.016)	0.456
	JB	0.636(0.056)	0.513(0.015)	0.158
0.75	TS-NLS	1.017(0.039)	0.461(0.011)	0.139
	TS-DR	1.003(0.160)	0.471(0.040)	0.646
	JB	0.971(0.051)	0.468(0.014)	0.225
0.90	TS-NLS	1.295(0.036)	0.426(0.010)	0.161
	TS-DR	1.411(0.077)	0.404(0.021)	0.732
	JB	1.245(0.054)	0.432(0.015)	0.380

Table 2.3: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as ZI Poisson distribution with proportion of zero-inflation $p(0) = 0.10$. Results based on a sample size $n = 250$.

τ	Method	$b_0(\tau)(s.e)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-0.281(0.082)	0.642(0.021)	0.215
	TS-DR	0.410(0.270)	0.466(0.075)	2.321
	JB	-0.927(0.504)	0.452(0.366)	25.640
0.25	TS-NLS	0.196(0.078)	0.574(0.020)	0.191
	TS-DR	0.664(0.107)	0.482(0.029)	1.703
	JB	-0.036(0.172)	0.614(0.045)	0.772
0.50	TS-NLS	0.649(0.067)	0.509(0.018)	0.186
	TS-DR	0.786(0.081)	0.492(0.022)	0.876
	JB	0.539(0.094)	0.528(0.026)	0.430
0.75	TS-NLS	1.016(0.062)	0.460(0.017)	0.210
	TS-DR	1.025(0.171)	0.470(0.043)	1.043
	JB	0.922(0.079)	0.475(0.022)	0.547
0.90	TS-NLS	1.295(0.056)	0.425(0.015)	0.244
	TS-DR	1.426(0.122)	0.403(0.034)	1.461
	JB	1.215(0.082)	0.436(0.023)	0.744

Table 2.4: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as ZI Poisson distribution with proportion of zero-inflation $p(0) = 0.10$. Results based on a sample size $n = 500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-0.281(0.059)	0.642(0.015)	0.176
	TS-DR	0.292(0.269)	0.499(0.073)	1.557
	JB	-0.922(0.407)	0.453(0.334)	25.956
0.25	TS-NLS	0.196(0.055)	0.574(0.014)	0.147
	TS-DR	0.639(0.107)	0.489(0.028)	1.605
	JB	-0.032(0.123)	0.614(0.032)	0.596
0.50	TS-NLS	0.649(0.049)	0.509(0.013)	0.135
	TS-DR	0.795(0.062)	0.489(0.017)	0.756
	JB	0.541(0.067)	0.527(0.018)	0.339
0.75	TS-NLS	1.015(0.043)	0.461(0.012)	0.146
	TS-DR	1.068(0.148)	0.459(0.037)	0.839
	JB	0.922(0.056)	0.475(0.016)	0.426
0.90	TS-NLS	1.295(0.040)	0.425(0.011)	0.169
	TS-DR	1.428(0.082)	0.402(0.023)	0.961
	JB	1.218(0.059)	0.436(0.016)	0.570

Table 2.5: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as ZI Poisson distribution with proportion of zero-inflation $p(0) = 0.45$. Results based on a sample size $n = 250$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-0.282(0.105)	0.643(0.027)	0.256
	TS-DR	0.656(0.107)	0.453(0.022)	4.495
	JB	-2.365(0.320)	0.046(0.107)	53.860
0.25	TS-NLS	0.191(0.102)	0.574(0.027)	0.252
	TS-DR	0.728(0.101)	0.492(0.028)	4.843
	JB	-1.363(0.267)	0.042(0.091)	76.421
0.50	TS-NLS	0.646(0.090)	0.510(0.024)	0.259
	TS-DR	0.827(0.122)	0.497(0.033)	2.678
	JB	-0.641(0.547)	0.628(0.303)	28.050
0.75	TS-NLS	1.015(0.079)	0.461(0.022)	0.288
	TS-DR	1.179(0.167)	0.442(0.043)	2.389
	JB	0.610(0.137)	0.522(0.037)	4.457
0.90	TS-NLS	1.296(0.075)	0.425(0.021)	0.351
	TS-DR	1.481(0.145)	0.399(0.041)	3.641
	JB	1.052(0.108)	0.459(0.030)	3.149

Table 2.6: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as ZI Poisson distribution with proportion of zero-inflation $p(0) = 0.10$. Results based on a sample size $n = 500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-0.280(0.075)	0.642(0.019)	0.198
	TS-DR	0.633(0.086)	0.461(0.019)	4.148
	JB	-2.290(0.201)	0.042(0.068)	53.585
0.25	TS-NLS	0.195(0.072)	0.574(0.019)	0.178
	TS-DR	0.733(0.076)	0.491(0.021)	4.721
	JB	-1.343(0.185)	0.040(0.064)	76.057
0.50	TS-NLS	0.648(0.062)	0.510(0.017)	0.169
	TS-DR	0.849(0.100)	0.491(0.026)	2.492
	JB	-0.672(0.404)	0.694(0.193)	20.453
0.75	TS-NLS	1.015(0.057)	0.461(0.016)	0.191
	TS-DR	1.224(0.108)	0.431(0.028)	2.059
	JB	0.615(0.096)	0.521(0.026)	4.189
0.90	TS-NLS	1.295(0.051)	0.426(0.014)	0.219
	TS-DR	1.483(0.090)	0.397(0.025)	2.193
	JB	1.054(0.075)	0.459(0.021)	2.896

Table 2.7: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as negative binomial distribution. Results based on a sample size $n = 250$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-1.538(0.230)	0.671(0.042)	0.240
	TS-DR	0.628(0.350)	-0.067(0.643)	3.158
	JB	-1.082(0.322)	0.593(0.099)	0.620
0.25	TS-NLS	-0.338(0.142)	0.556(0.039)	0.352
	TS-DR	0.515(0.259)	0.342(0.091)	1.905
	JB	-0.232(0.211)	0.540(0.067)	0.651
0.50	TS-NLS	0.466(0.120)	0.510(0.037)	0.615
	TS-DR	0.707(0.120)	0.456(0.043)	1.335
	JB	0.448(0.156)	0.515(0.050)	0.974
0.75	TS-NLS	1.055(0.109)	0.489(0.036)	1.375
	TS-DR	0.979(0.221)	0.521(0.069)	4.729
	JB	1.009(0.135)	0.498(0.044)	2.128
0.90	TS-NLS	1.482(0.111)	0.478(0.036)	3.148
	TS-DR	1.502(0.266)	0.481(0.082)	10.460
	JB	1.443(0.141)	0.484(0.047)	5.114

Table 2.8: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as negative binomial distribution. Results based on a sample size $n = 500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-1.542(0.163)	0.671(0.030)	0.176
	TS-DR	0.539(0.441)	-0.039(0.668)	2.985
	JB	-1.081(0.229)	0.594(0.071)	0.448
0.25	TS-NLS	-0.332(0.100)	0.554(0.027)	0.211
	TS-DR	0.374(0.285)	0.385(0.088)	1.319
	JB	-0.224(0.146)	0.539(0.047)	0.400
0.50	TS-NLS	0.468(0.086)	0.510(0.027)	0.357
	TS-DR	0.701(0.089)	0.458(0.032)	0.861
	JB	0.450(0.111)	0.515(0.036)	0.551
0.75	TS-NLS	1.056(0.078)	0.490(0.026)	0.735
	TS-DR	1.062(0.183)	0.496(0.057)	2.574
	JB	1.007(0.097)	0.499(0.032)	1.112
0.90	TS-NLS	1.484(0.079)	0.478(0.026)	1.649
	TS-DR	1.530(0.203)	0.473(0.062)	5.509
	JB	1.447(0.101)	0.483(0.033)	2.701

Table 2.9: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as ZINB distribution with proportion of zero-inflation $p(0) = 0.10$. Results based on a sample size $n = 250$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-1.537(0.320)	0.671(0.052)	0.326
	TS-DR	0.380(0.513)	0.268(0.224)	4.759
	JB	-1.468(0.380)	0.351(0.189)	2.881
0.25	TS-NLS	-0.339(0.190)	0.556(0.045)	0.471
	TS-DR	0.582(0.229)	0.373(0.076)	3.328
	JB	-0.660(0.306)	0.566(0.101)	1.930
0.50	TS-NLS	0.467(0.146)	0.510(0.042)	0.783
	TS-DR	0.727(0.150)	0.475(0.048)	3.156
	JB	0.285(0.188)	0.527(0.060)	1.947
0.75	TS-NLS	1.059(0.131)	0.488(0.042)	1.652
	TS-DR	1.040(0.260)	0.515(0.079)	7.173
	JB	0.933(0.147)	0.502(0.048)	3.129
0.90	TS-NLS	1.485(0.127)	0.476(0.041)	3.562
	TS-DR	1.543(0.270)	0.476(0.083)	11.786
	JB	1.402(0.152)	0.484(0.050)	6.508

Table 2.10: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as ZINB distribution with proportion of zero-inflation $p(0) = 0.10$. Results based on a sample size $n = 500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-1.544(0.221)	0.671(0.036)	0.213
	TS-DR	0.194(0.564)	0.326(0.211)	4.178
	JB	-1.427(0.270)	0.332(0.140)	3.008
0.25	TS-NLS	-0.336(0.137)	0.555(0.032)	0.263
	TS-DR	0.506(0.230)	0.396(0.073)	2.826
	JB	-0.659(0.221)	0.568(0.072)	1.630
0.50	TS-NLS	0.469(0.102)	0.510(0.030)	0.444
	TS-DR	0.718(0.112)	0.479(0.036)	2.497
	JB	0.290(0.131)	0.526(0.042)	1.513
0.75	TS-NLS	1.056(0.092)	0.489(0.029)	0.853
	TS-DR	1.120(0.201)	0.493(0.060)	4.277
	JB	0.934(0.105)	0.502(0.034)	2.180
0.90	TS-NLS	1.487(0.088)	0.477(0.029)	1.887
	TS-DR	1.562(0.203)	0.472(0.062)	7.182
	JB	1.406(0.105)	0.485(0.035)	3.937

Table 2.11: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as ZINB distribution with proportion of zero-inflation $p(0) = 0.45$. Results based on a sample size $n = 250$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-1.556(1.048)	0.672(0.068)	0.474
	TS-DR	0.566(0.266)	0.415(0.087)	23.911
	JB	-2.700(0.367)	0.118(0.116)	5.316
0.25	TS-NLS	-0.389(1.580)	0.556(0.058)	0.768
	TS-DR	0.691(0.191)	0.446(0.067)	18.361
	JB	-1.637(0.270)	0.101(0.092)	21.505
0.50	TS-NLS	0.462(0.198)	0.511(0.055)	1.276
	TS-DR	0.784(0.214)	0.516(0.070)	17.744
	JB	-1.176(0.497)	0.479(0.289)	50.175
0.75	TS-NLS	1.051(0.173)	0.489(0.055)	2.648
	TS-DR	1.172(0.323)	0.512(0.098)	22.463
	JB	0.440(0.270)	0.534(0.085)	29.723
0.90	TS-NLS	1.477(0.161)	0.478(0.053)	5.869
	TS-DR	1.629(0.278)	0.475(0.088)	28.203
	JB	1.157(0.196)	0.497(0.064)	28.446

Table 2.12: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as ZINB distribution with proportion of zero-inflation $p(0) = 0.45$. Results based on a sample size $n = 500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.10	TS-NLS	-1.552(0.743)	0.671(0.047)	0.288
	TS-DR	0.474(0.251)	0.443(0.079)	23.701
	JB	-2.573(0.211)	0.102(0.070)	5.313
0.25	TS-NLS	-0.338(0.172)	0.556(0.040)	0.396
	TS-DR	0.679(0.141)	0.451(0.046)	17.454
	JB	-1.620(0.183)	0.099(0.063)	21.501
0.50	TS-NLS	0.466(0.139)	0.510(0.039)	0.673
	TS-DR	0.795(0.164)	0.514(0.052)	15.953
	JB	-1.218(0.342)	0.507(0.224)	50.346
0.75	TS-NLS	1.055(0.119)	0.489(0.038)	1.371
	TS-DR	1.256(0.236)	0.490(0.072)	16.303
	JB	0.453(0.180)	0.532(0.058)	28.734
0.90	TS-NLS	1.483(0.117)	0.478(0.038)	3.118
	TS-DR	1.644(0.202)	0.471(0.063)	18.312
	JB	1.164(0.140)	0.496(0.046)	26.251

In order to explore the properties of the estimators, another set of simulations were performed with sample size $n = 1000000$. This enables researchers to check the large-sample performance of the estimators. Also, this enables researchers to obtain the pseudo-true $\beta(\tau)$. This simulation-based strategy to calculate the pseudo-true parameter values is done so as defined in the context of model selection; see Sawa (1978) and Vuong (1989).

Table 2.13: Pseudo-true parameter values for QR coefficients. Data generated from Poisson distribution with a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	MISE
0.10	TS-NLS	-0.279	0.642	0.14
	TS-DR	-0.153	0.588	0.74
	JB	-0.238	0.643	0.26
0.25	TS-NLS	0.195	0.574	0.10
	TS-DR	0.379	0.541	0.37
	JB	0.238	0.570	0.16
0.50	TS-NLS	0.648	0.510	0.08
	TS-DR	0.814	0.477	0.34
	JB	0.636	0.513	0.08
0.75	TS-NLS	1.016	0.461	0.08
	TS-DR	1.151	0.432	0.30
	JB	0.969	0.468	0.14
0.90	TS-NLS	1.295	0.425	0.10
	TS-DR	1.407	0.401	0.30
	JB	1.248	0.432	0.21

Table 2.14: Pseudo-true parameter values for QR coefficients. Data generated from Poisson distribution with a proportion of zero-inflation $p(0) = 0.10$. Results computed based on a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	MISE
	TS-NLS	-0.277	0.642	0.14
0.10	TS-DR	1.336	0.008	26.56
	JB	-0.923	0.425	34.06
	TS-NLS	0.198	0.574	0.10
0.25	TS-DR	0.375	0.551	0.78
	JB	-0.030	0.615	0.42
	TS-NLS	0.650	0.510	0.08
0.50	TS-DR	0.823	0.478	0.43
	JB	0.542	0.528	0.23
	TS-NLS	1.017	0.461	0.08
0.75	TS-DR	1.165	0.432	0.46
	JB	0.926	0.474	0.32
	TS-NLS	1.297	0.425	0.10
0.90	TS-DR	1.421	0.402	0.50
	JB	1.221	0.435	0.41

Table 2.15: Pseudo-true parameter values for QR coefficients. Data generated from Poisson distribution with a proportion of zero-inflation $p(0) = 0.45$. Results computed based on a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	MISE
	TS-NLS	-0.279	0.642	0.14
0.10	TS-DR	0.569	0.478	3.43
	JB	-2.244	0.038	53.38
	TS-NLS	0.195	0.574	0.10
0.25	TS-DR	0.731	0.492	4.58
	JB	-1.328	0.037	75.69
	TS-NLS	0.648	0.510	0.08
0.50	TS-DR	1.009	0.452	2.73
	JB	-0.708	0.742	14.25
	TS-NLS	1.016	0.461	0.08
0.75	TS-DR	1.249	0.423	1.63
	JB	0.617	0.521	3.96
	TS-NLS	1.295	0.425	0.10
0.90	TS-DR	1.457	0.402	1.41
	JB	1.056	0.458	2.72

Table 2.16: Pseudo-true parameter values for QR coefficients. Data generated from Poisson distribution with two predictors for the mean model specification. Results computed based on a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	MISE
	TS-NLS	-0.376	0.669	-0.467	0.14
0.10	TS-DR	0.691	-0.034	-0.059	8.15
	JB	-0.491	0.731	-0.501	0.34
	TS-NLS	0.142	0.589	-0.410	0.12
0.25	TS-DR	0.859	0.038	-0.074	10.43
	JB	0.147	0.607	-0.425	0.21
	TS-NLS	0.635	0.514	-0.360	0.09
0.50	TS-DR	1.103	0.045	0.004	14.51
	JB	0.598	0.530	-0.372	0.11
	TS-NLS	1.043	0.453	-0.320	0.08
0.75	TS-DR	1.226	0.400	-0.282	0.37
	JB	0.971	0.468	-0.328	0.13
	TS-NLS	1.349	0.410	-0.290	0.11
0.90	TS-DR	1.495	0.366	-0.260	0.39
	JB	1.277	0.419	-0.295	0.22

Table 2.17: Pseudo-true parameter values for QR coefficients. Data generated from ZIP distribution with a proportion of zero-inflation $p(0) = 0.10$. Two predictors were used for the mean parameters. Results computed based on a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	MISE
	TS-NLS	-0.380	0.669	-0.466	0.14
0.10	TS-DR	0.703	-0.126	0.013	9.42
	JB	-1.081	0.599	-0.429	4.65
	TS-NLS	0.138	0.589	-0.409	0.12
0.25	TS-DR	0.745	0.058	-0.017	10.42
	JB	-0.139	0.654	-0.452	0.22
	TS-NLS	0.633	0.513	-0.358	0.09
0.50	TS-DR	1.215	0.011	0.025	15.55
	JB	0.471	0.556	-0.388	0.18
	TS-NLS	1.041	0.453	-0.318	0.08
0.75	TS-DR	1.245	0.401	-0.285	0.41
	JB	0.915	0.478	-0.335	0.24
	TS-NLS	1.348	0.409	-0.289	0.11
0.90	TS-DR	1.522	0.365	-0.264	0.45
	JB	1.244	0.424	-0.298	0.35

Table 2.18: Pseudo-true parameter values for QR coefficients. Data generated from ZIP distribution with a proportion of zero-inflation $p(0) = 0.45$. Two predictors were used for the mean parameters. Results computed based on a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	MISE
0.10	TS-NLS	-0.382	0.671	-0.467	0.14
	TS-DR	0.769	0.055	-0.026	7.27
	JB	-2.535	0.177	-0.122	11.56
0.25	TS-NLS	0.136	0.591	-0.410	0.12
	TS-DR	1.179	0.000	0.006	11.59
	JB	-1.613	0.176	-0.124	17.88
0.50	TS-NLS	0.630	0.515	-0.359	0.09
	TS-DR	1.437	-0.000	-0.001	15.25
	JB	-0.555	0.690	-0.471	6.83
0.75	TS-NLS	1.038	0.454	-0.319	0.08
	TS-DR	1.308	0.407	-0.288	1.20
	JB	0.523	0.560	-0.391	2.00
0.90	TS-NLS	1.345	0.411	-0.290	0.11
	TS-DR	1.563	0.369	-0.261	1.18
	JB	1.060	0.458	-0.322	1.56

Table 2.19: Pseudo-true parameter values for QR coefficients. Data generated from negative binomial distribution with a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	MISE
0.10	TS-NLS	-0.777	0.613	0.14
	TS-DR	0.774	-0.112	10.14
	JB	-0.667	0.608	0.29
0.25	TS-NLS	-0.070	0.557	0.09
	TS-DR	0.179	0.511	0.33
	JB	0.010	0.546	0.15
0.50	TS-NLS	0.561	0.510	0.08
	TS-DR	0.735	0.475	0.30
	JB	0.550	0.513	0.08
0.75	TS-NLS	1.039	0.484	0.09
	TS-DR	1.184	0.451	0.33
	JB	1.002	0.489	0.14
0.90	TS-NLS	1.397	0.468	0.11
	TS-DR	1.523	0.437	0.45
	JB	1.364	0.471	0.25

Table 2.20: Pseudo-true parameter values for QR coefficients. Data generated from negative binomial distribution with a proportion of zero-inflation $p(0) = 0.10$. Results computed based on a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	MISE
0.10	TS-NLS	-0.780	0.613	0.14
	TS-DR	0.881	-0.023	7.90
	JB	-1.130	0.318	9.97
0.25	TS-NLS	-0.074	0.558	0.09
	TS-DR	0.140	0.553	2.00
	JB	-0.360	0.587	1.06
0.50	TS-NLS	0.558	0.510	0.08
	TS-DR	0.734	0.491	1.10
	JB	0.417	0.526	0.66
0.75	TS-NLS	1.037	0.484	0.08
	TS-DR	1.192	0.462	1.13
	JB	0.936	0.995	0.77
0.90	TS-NLS	1.395	0.468	0.11
	TS-DR	1.541	0.443	1.11
	JB	1.322	0.474	0.99

Table 2.21: Pseudo-true parameter values for QR coefficients. Data generated from negative binomial distribution with a proportion of zero-inflation $p(0) = 0.45$. Results computed based on a sample size of $n = 1,000,000$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	MISE
0.10	TS-NLS	-0.777	0.613	0.14
	TS-DR	0.685	0.317	4.01
	JB	-2.385	0.072	15.08
0.25	TS-NLS	-0.070	0.557	0.09
	TS-DR	0.632	0.492	12.62
	JB	-1.467	0.071	37.88
0.50	TS-NLS	0.561	0.510	0.08
	TS-DR	1.013	0.460	8.43
	JB	-1.118	0.6827	38.16
0.75	TS-NLS	1.039	0.484	0.09
	TS-DR	1.332	0.449	6.15
	JB	0.540	0.528	15.82
0.90	TS-NLS	1.397	0.468	0.11
	TS-DR	1.607	0.440	4.60
	JB	1.119	0.488	12.05

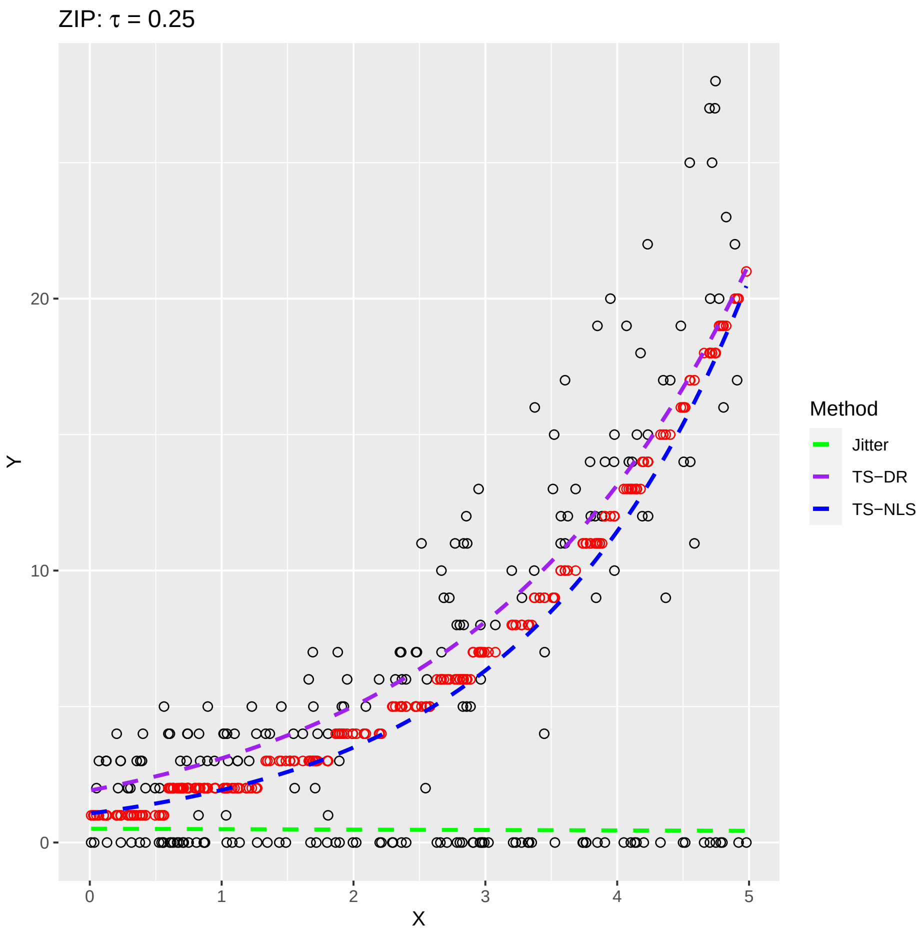


Figure 2.4: Comparison of different methods when zero-inflation exists; data generated from ZIP distribution with proportion of zero-inflation $p(0) = 0.10$. Red dots are the true conditional quantiles at $\tau = 0.25$.

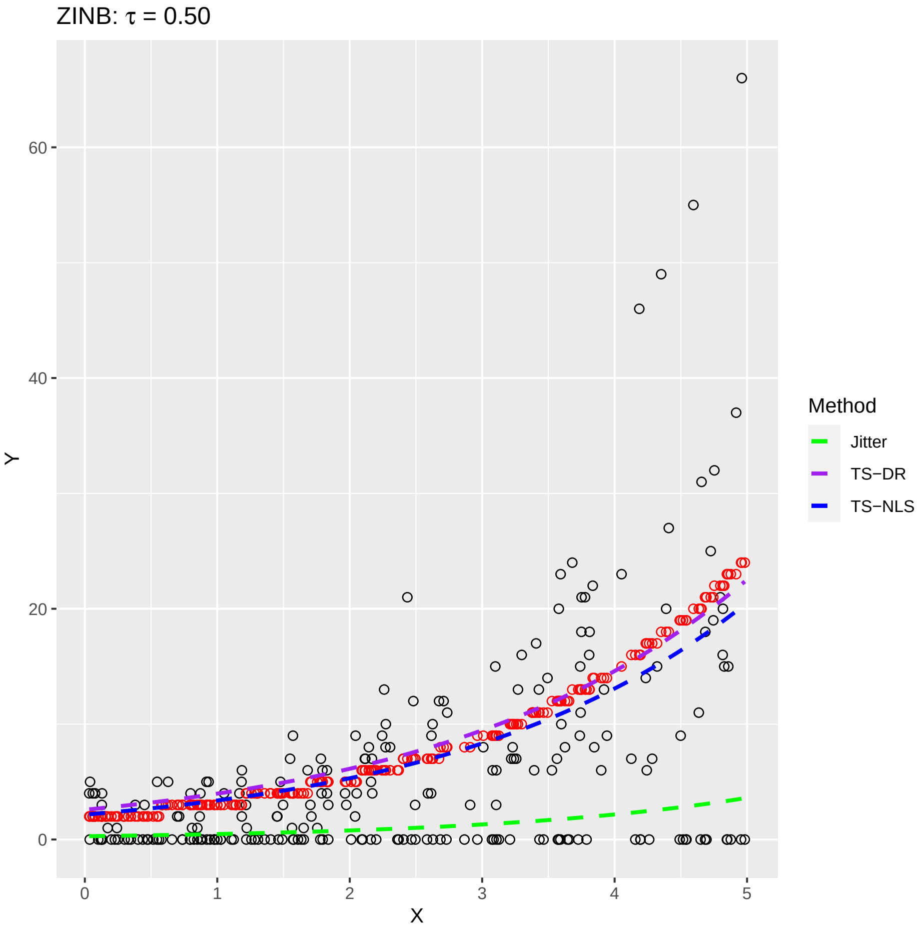


Figure 2.5: Comparison of different methods when zero-inflation exists; data generated from ZINB distribution with proportion of zero-inflation $p(0) = 0.45$. Red dots are the true conditional quantiles at $\tau = 0.50$.

2.6 Data Analysis

In 2008, the state of Oregon initiated a Medicaid expansion program to provide health care coverage for its low-income, uninsured residents. Individuals were first required to sign up for a lottery during a five-week window; the state then conducted eight drawings from the lottery list. The winners of the lottery were given the opportunity to apply for the Oregon Health Plan (OHP) Standard for themselves and any listed (or unlisted) household members. If these selected individuals submitted the application following the instructions, and they met the eligibility criteria set by the state, they would be enrolled in OHP Standard. Details about the Oregon health insurance experiment can be found in Finkelstein et al. (2012).

From a statistical point of view, the Oregon health insurance experiment represents a large-scale randomized experiment in which the lottery mechanism corresponds to a random assignment to treatments. Multiple data sets have been collected since the experiment, and many have been analyzed by researchers from different perspectives (Finkelstein et al., 2012; Baicker et al., 2013, 2014).

In this section, we analyzed one specific data source from the Oregon health insurance experiment. This data set was first introduced as the Survey Data in Section V of Finkelstein et al. (2012). Chernozhukov et al. (2018) studied this data set after excluding subjects with incomplete information in selected variables. The resulting subset consists of $13,173$ observations. The response variable is the count of outpatient visits during a six-month period and $5,027$ of all the observations are zero counts. That is, about 38% of the counts are zero. A histogram of the response variable is provided in Figure (2.6) for illustration. The independent variable of interest is an indicator variable for whether one household won the lottery (Treatment) or not (Control); hence, the corresponding regression coefficient can be interpreted as the intent-to-treat (ITT) effect, the effect of being able to apply for OHP Standard on the health care utilization. Other covariates include the number of prescription medications currently taking (truncated at 2×99^{th} percentile), the indicator variables for the number of household individuals, the indicator variables for the survey waves, and their interactions.

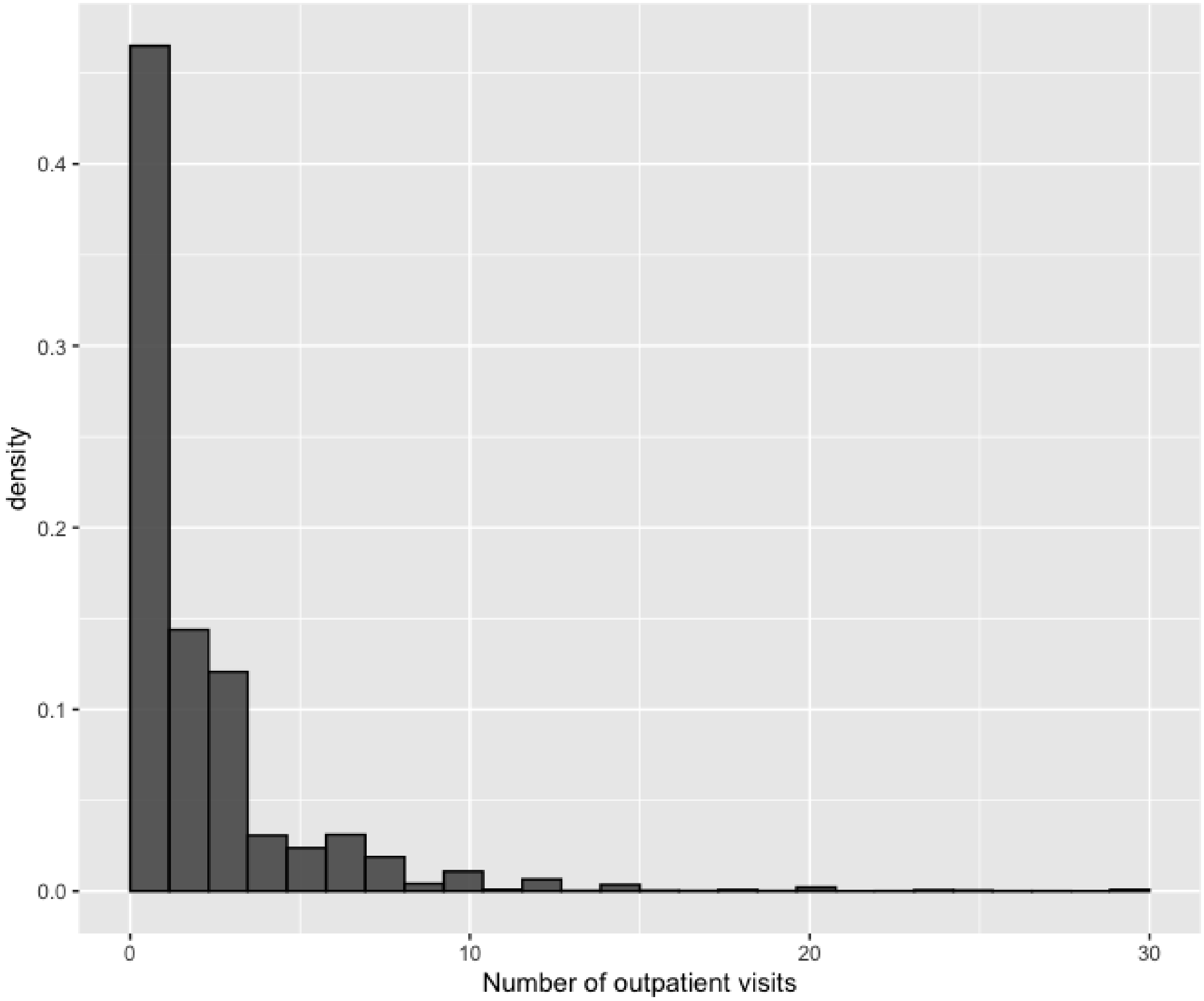


Figure 2.6: Number of visits to physicians in the Oregon Health Insurance Experiment data.

In the real life, there are two mechanisms that lead to a zero count in the number of outpatient visits during a specific period. The first mechanism, which corresponds to the *perfect state* as in Lambert (1992), is when people are healthy during the period so that they have no need to visit hospitals at all. In this case, a structural zero is simply the only possibility if researchers record the number of outpatient visits. The second mechanism, which corresponds to the *imperfect state* in the same paper, happens when patients contract diseases and need to visit hospitals. Depending on the illness and the economic consideration, some patients might not go to the hospitals while others have multiple visits. Under this circumstance, a random zero count could be observed but a positive integer is also possible.

In the Oregon health insurance experiment, researchers wanted to estimate the ITT effect of winning the lottery on health care utilization. Based on the literature and real-life experience, we believe it is a decent assumption that not all the participants experienced illness during the 6-month period. Hence, these participants were in the perfect state and had no intent to visit hospitals. To analyze the ITT effect as specified, it is necessary to distinguish healthy participants from participants who need to visit hospitals. As a result, the application of ZI models is plausible in the analysis.

We began the analyses with the conditional mean structure. Finkelstein et al. (2012) considered a linear model with the following specification:

$$y_i = \beta_0 + \beta_1 D_i + \mathbf{X}_i^{'}\boldsymbol{\beta} + \epsilon_i, \tag{2.29}$$

where D_i is an indicator variable for whether the household of subject i was selected by the lottery. Thus, the regression coefficient β_1 is the coefficient of interest and estimates the ITT effect. $\mathbf{X}$ includes the indicator variables for the number of individuals in the household, the indicator variables for survey wave and the interaction of these indicator variables. In this paper, we first considered a similar model specification with the same set of predictors. The only difference is that we also added one numerical predictor, the number of prescription medications currently taking, into our model specification. This model specification will be denoted as the *full model* thereafter.

Based on the existing analyses (Finkelstein et al., 2012; Baicker et al., 2013, 2014), we also considered a simpler model specification where $\mathbf{X}$ only includes the significant predictors; hence, this specification keeps the indicator variables for the number of individuals in the household and the number of prescription medications currently taking. The simpler specification, denoted as the *reduced model* thereafter, is parsimonious and helps the researchers focus on the illustration of the proposed method. However, it also turns out that this specification provides a better fit to the data compared to the *full model*.

Given the fact that the responses are counts and certain number of observations are zero counts, we considered the four aforementioned count regression models: Poisson regression model, ZIP regression model, negative binomial regression model and ZINB regression model. The linear regression model was included for the purpose of comparison. Model comparisons via the BIC values are provided in Table (2.22). From Table (2.22), it is obvious that the four count regression models provide better fit than the linear regression model; it is also obvious that the *reduced model* fits the data better than the *full model*. Thus, the following analyses focuses on the *reduced model* specification with count regression models.

Table 2.22: Model comparisons via BIC values with degree of freedoms (df) in the parentheses.

Model	Full(df)	Reduced(df)
Linear Regression	$63678.56(20)$	$63561.00(6)$
Poisson Regression	$58280.13(19)$	$58207.00(5)$
ZIP Regression	$50972.68(21)$	$50902.23(7)$
NB Regression	$47979.72(20)$	$47863.83(6)$
ZINB Regression	$45412.40(22)$	$45301.08(8)$

In the first step, we modeled the mean structure of the data by four different count regression models: Poisson regression model, negative binomial regression model, ZIP regression model and ZINB regression model. This helps researchers explore the overall structure of the data, and also provides the starting point for the three-step routine in QR modeling. The randomized quantile residuals under each model is used for GOF assessment.

A comparison reveals huge differences among models. Clearly, Poisson regression model shows the worst fit, due to the fact that it fails to capture the presence of zero-inflation and certain large values. ZIP regression shows some improvements over Poisson regression, but still fails to tackle the overdispersion. On the other hand, both negative binomial regression and ZINB regression provide satisfactory fit to the data, as indicated by the randomized quantile residual plot in Figure 2.7.

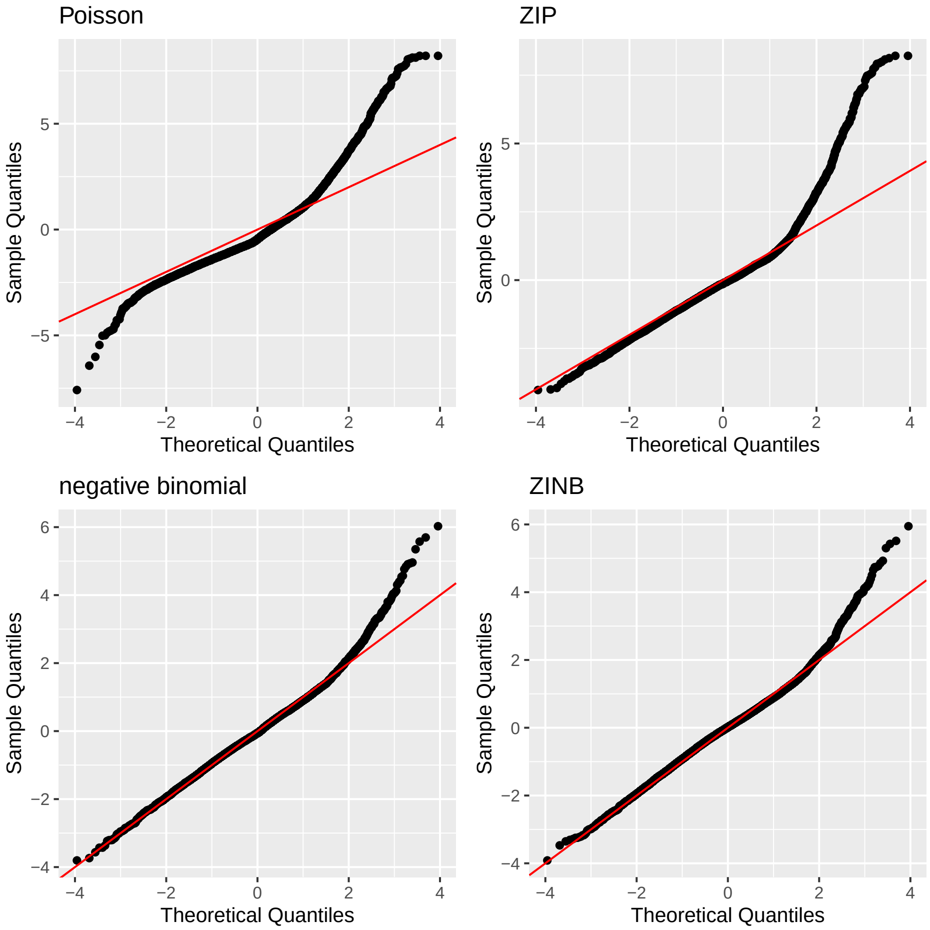


Figure 2.7: Randomized quantile residuals of fitted models. First row shows results for Poisson regression (left) and negative binomial regression (right); second row shows results for ZIP regression (left) and ZINB (right).

To test for the presence of zero-inflation, we conducted a boundary-corrected LR test. Results for three tests are reported in Table 2.23. All the tests show evidence for the presence of zero-inflation, and ZINB regression gives the best fit based on these results.

Table 2.23: Results of the boundary likelihood ratio test. All p-values are significant at 0.001

Models tested	Test statistic	Results
ZIP *versus* Poisson	7323.8	ZIP
ZINB *versus* negative binomial	2581.8	ZINB
ZINB *versus* ZIP	5610.6	ZINB

Table 2.24: Estimated regression coefficients, $\hat{\beta}$ and $\hat{\beta}^{\tau}$, for the Oregon Health Insurance Experiment data.

Variables	**Lottery**	Prescription	Household Size $(= 2)$	Household Size (≥ 3)
Mean (ZINB)	0.084 (0.020)	0.113 (0.004)	-0.095 (0.022)	-0.515 (0.251)
$\tau = 0.30$	0.126 (0.028)	0.121 (0.005)	-0.141 (0.039)	-0.794 (0.241)
$\tau = 0.40$	0.106 (0.019)	0.125 (0.004)	-0.096 (0.019)	-0.754 (0.253)
$\tau = 0.50$	0.092 (0.042)	0.125 (0.005)	-0.099 (0.043)	-0.420 (0.205)
$\tau = 0.60$	0.036 (0.031)	0.112 (0.004)	-0.060 (0.029)	-0.533 (0.192)
$\tau = 0.70$	0.111 (0.036)	0.109 (0.005)	-0.120 (0.030)	-0.473 (0.233)
$\tau = 0.80$	0.094 (0.022)	0.109 (0.004)	-0.089 (0.019)	-0.507 (0.197)
$\tau = 0.90$	0.091 (0.022)	0.104 (0.004)	-0.096 (0.020)	-0.471 (0.191)

The next part models the conditional quantiles of the count process. Table 2.24 presents results for the conditional mean effects and quantile effects corresponding to ITT(treatment) effect, health status and demographics. The first row presents point estimates for the mean parameters, and the following rows show results for the quantile parameters. The table also reports SE of the estimator obtained by the multiplier bootstrap. 200 bootstrap samples are used to compute the SE.

The point estimates corresponding to the first step are in row 1 of Table 2.24. As supported by earlier evidence, the ZINB regression model provides a better fit to the data, due to the fact that the data shows both zero-inflation and over-dispersion. The positive sign of the estimated coefficients for the variable Lottery indicates a positive ITT effect. That is, participants in the treatment group have higher health care utilization. This is consistent with standard public health results in the literature. For instance, Finkelstein et al. (2012) found the same conclusion regarding the ITT effect. When we examine the ITT effects across quantiles, we notice some variability compared to that at the mean level. The estimated value is greater at the lower quantile ($\tau = 0.30$), then drops gradually as the quantile increases up to ($\tau = 0.60$). The value then increases as the quantile keeps increasing. When ($\tau = 0.70$), it is similar to that at ($\tau = 0.30$). Finally, the estimates seems to drop back and converge to the value at the mean level. It is also noticeable that the mean ITT estimates is quantitatively similar to the estimated effects at the median ($\tau = 0.50$) and the top quantiles ($\tau \geq 0.80$). In summary, these results reveal the difference of distributional effects, depending on the conditional quantiles at which the participants are located. At the same time, the finding also suggest that the mean effect captures an incomplete description of the variable.

Results about other predictors also provide important findings. The variable Prescription, the number of prescription medicines one is currently taking, can be interpreted as an indicator of health and is directly related to the health care utilization. The sign is positive and significant at every levels, which agrees with our expectation. In the real life, one would expect the less healthy people to have more prescriptions, and to have higher numbers of outpatient visits if they can. At the same time, we notice that the effects of Prescription are greater at the lower quantiles.

The signs of the indicators for household size are all negative, suggesting that a single person is more likely to have higher number of outpatient visits. This makes sense as family members can help take care of each other, while a single person has to visit doctors and hospitals in case of illness.

Figure 2.8 and 2.9 provide visual illustrations of the estimated mean effects (dashed lines) and quantile effects (continuous lines) in Table 2.24. In addition, we would like to show the importance of accommodating for the zero-inflation in the data. Hence, we also report estimates obtained by the jittering approach (Machado and Santos Silva, 2005). These figure show some interesting differences. First, we find that the ITT effect associated with Lottery varies differently across quantiles, particularly among those at lower and around the median. Secondly, the estimates for the effect of prescription medication show huge gap between the two methods. These findings suggest that when the data shows zero-inflation, the corresponding analyses should accommodate the effects of excess zeros.

2.7 Discussion

Quantiles provide a more comprehensive description of the conditional distribution, yet classic QR is not feasible for data with discrete responses. At the same time, the presence of zero-inflation in a count data makes inference even more complicated. In this paper, we proposed a three-step approach for modeling the conditional quantiles of count data and ZI data . Our approach is essentially parametric, where all existing methods for regression

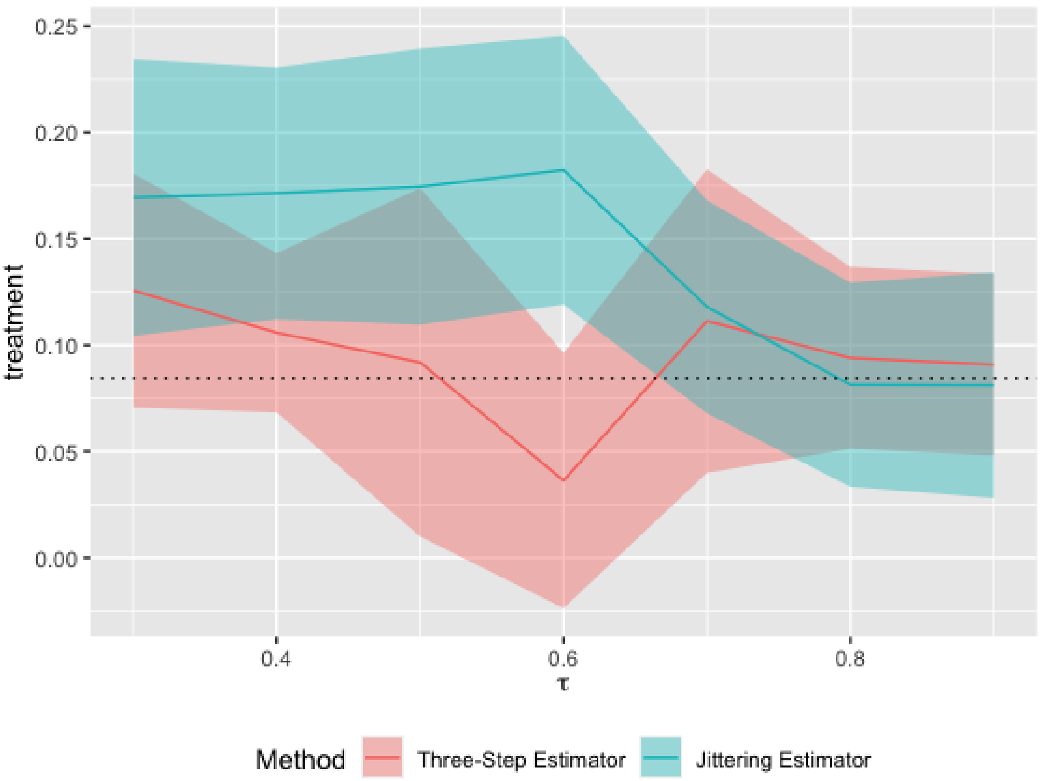


Figure 2.8: Estimated ITT effect (regression coefficient for treatment) in the Oregon Health Insurance Experiment data. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

can be applied directly. This also includes extension to semi-parametric regression and non-parametric regression. When the data shows evidence for a particular distribution, our method gives better results by making use of that information. In particular, the extension of our method to ZI data demonstrates an efficacious modeling strategy.

Theoretically, a parametric model is less robust to mis-specification of which distribution to use. However, as stated earlier, in the first step of our approach the estimation of the conditional mean function can be carried out by semi-parametric or non-parametric methods. This gives hope to provide more robust extensions of our method, and will be one of our research topics in the future.

Our approach is based on interpolation of the discrete distribution for the count model under consideration. This approach was thoroughly introduced in Ilienko (2013b) and

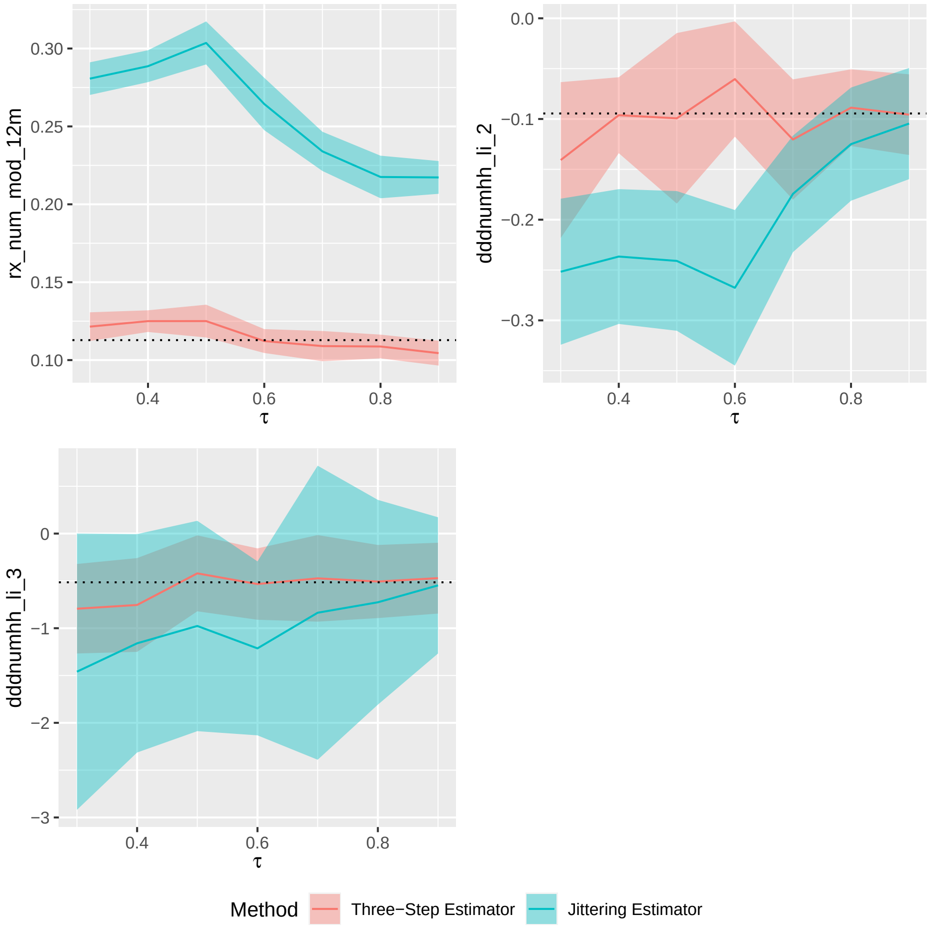


Figure 2.9: Estimated regression coefficient for covariates in the Oregon Health Insurance Experiment data. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

further developed in Padellini and Rue (2019a), with the latter extending this approximation to perform quantile regression for discrete data. An important distinction of our method from Padellini and Rue (2019a) is that we first consider estimation of the conditional mean rather than the conditional quantiles. This flexibility allows us to employ existing methods for consistent estimation of a class of models. In a ZI setting, a convenient yet powerful class of models for the first step are ZI GLMMs. This broad class of models enables researchers to extensively explore reasonable distributions to for the conditional mean structure. Computationally, modeling can be conveniently performed using the `R` package `glmmTMB` (Brooks et al., 2017).

The interpolation technique introduces a continuous counterpart for the discrete distribution of choice. Given that the requirement of continuous distributions is usually the first regularity condition for deriving the asymptotic distribution, the use of continuous Poisson or continuous negative binomial distribution gives researchers new directions for inference about the regression parameters.

We have seen that the extension of our method to ZI problems provides a novel modeling strategy. This is highlighted by analyzing data from the Oregon Health Insurance Experiment. While these data have been analyzed in the literature based on mean regression models, we extended the analyses with a through examination of quantile effects while capturing zero-inflation. Our analysis provides a more comprehensive view that can help public health policymakers and researchers understand how certain policies affect the participants differently. Overall, the empirical results obtained for this data analysis, combined with the extensive simulation results, suggest the benefit of our novel techniques to model quantile effects when ZI count responses occur.

2.8 Appendix

Interpolation for Poisson Distribution

Proof. Suppose $X \sim poisson(\lambda)$, then:

$$\begin{aligned}
F_X(x) &= P(X \le x), x \ge 0 \\
&= P(X \le \lfloor x \rfloor), \text{since X only takes non-negative integer values.} \\
&= P(Y > \lambda), Y \sim Gamma(\lfloor x \rfloor + 1, 1) \\
&= \int_{\lambda}^{\infty} \frac{1}{\Gamma(\lfloor x \rfloor + 1)} \cdot y^{\lfloor x \rfloor} \cdot e^{-y} dy \\
&= \frac{\int_{\lambda}^{\infty} y^{\lfloor x \rfloor} \cdot e^{-y} dy}{\Gamma(\lfloor x \rfloor + 1)}, \text{numerator is the definition of incomplete gamma function.} \\
&= \frac{\Gamma(\lfloor x \rfloor + 1, \lambda)}{\Gamma(\lfloor x \rfloor + 1)}
\end{aligned} \tag{2.30}$$

Then, the continuous counterpart of X is obtained by removing the floor function. Suppose $X \sim poisson(\lambda)$ and X' is the continuous counterpart of X by interpolation, then:

$$F_{X'}(x) = \frac{\Gamma(x+1,\lambda)}{\Gamma(x+1)}, x \geq 0 \tag{2.31}$$

For non-negative integer x, have:

$$\begin{aligned}
P(\lceil X^{'} \rceil = x) &= P(X^{'} \in (x-1, x]) \\
&= F_{X'}(x) - F_{X'}(x-1) \\
&= \frac{\Gamma(x+1,\lambda)}{\Gamma(x+1)} - \frac{\Gamma(x,\lambda)}{\Gamma(x)} \\
&= \frac{\Gamma(x+1,\lambda)}{\Gamma(x+1)} - \frac{x \cdot \Gamma(x,\lambda)}{\Gamma(x+1)} \\
&= \frac{\Gamma(x+1,\lambda) - x \cdot \Gamma(x,\lambda)}{\Gamma(x+1)}
\end{aligned} \tag{2.32}$$

where

$$\Gamma(x+1,\lambda) - x \cdot \Gamma(x,\lambda) = \int_\lambda^\infty e^{-s} \cdot s^x ds - x \cdot \int_\lambda^\infty e^{-s} \cdot s^{x-1} ds, \text{by definition} \tag{2.33}$$

Using integration by parts with $u = s^x$ and $dv = e^{-s}$,

$$\begin{aligned}
\int_\lambda^\infty e^{-s} \cdot s^x ds &= s^x \cdot (-e^{-s}) \mid_{s=\lambda}^{\infty} + \int_\lambda^\infty x \cdot e^{-s} \cdot s^{x-1} ds \\
&= \lambda^x \cdot e^{-\lambda} + \int_\lambda^\infty x \cdot e^{-s} \cdot s^{x-1} ds
\end{aligned}$$

Hence,

$$\begin{aligned}
\Gamma(x+1,\lambda) - x \cdot \Gamma(x,\lambda) &= \lambda^x \cdot e^{-\lambda} + \int_\lambda^\infty x \cdot e^{-s} \cdot s^{x-1} ds - \int_\lambda^\infty x \cdot e^{-s} \cdot s^{x-1} ds \\
&= \lambda^x \cdot e^{-\lambda} \\
\Rightarrow P(\lceil X^{'} \rceil = x) &= \frac{\Gamma(x+1,\lambda) - x \cdot \Gamma(x,\lambda)}{\Gamma(x+1)} \\
&= \frac{\lambda^x \cdot e^{-\lambda}}{x!} \\
&= P(X = x), \text{ for non-negative integer x.}
\end{aligned}$$

□

Interpolation for Negative Binomial Distribution

Proof. Suppose Z is the number of successes to get r failures in a series of independent Bernoulli Trials with a success probability equal p. By definition, $Z \sim Negative\ Binomial(r, p)$ with pmf:

$$f_Z(z) = \binom{z+r-1}{z} \cdot p^z \cdot (1-p)^r \tag{2.34}$$

and CDF:

$$\begin{aligned}
F_Z(\lfloor z \rfloor) &= P(Z \leq \lfloor z \rfloor), \text{for z ¿ 0} \\
&= 1 - P(Z > \lfloor z \rfloor) \\
&= 1 - P(Z \geq \lfloor z \rfloor + 1) \\
&= 1 - P(Z \geq k), \text{where k = } \lfloor z \rfloor \text{ + 1} \\
&= 1 - \sum_{z=k}^{\infty} \binom{z+r-1}{z} \cdot p^z \cdot (1-p)^r \\
&= 1 - \sum_{z=k}^{\infty} \binom{z+r-1}{z} \cdot \frac{t^z}{(1+t)^{z+r}}, \text{where p = } \frac{t}{1+t}
\end{aligned}$$

Taking derivative w.r.t t yields,

$$\begin{aligned}
-\frac{\partial F_Z(\lfloor z \rfloor)}{\partial t} &= \sum_{z=k}^{\infty} \binom{z+r-1}{z} \cdot \left[\frac{z \cdot t^{z-1}}{(1+t)^{z+r}} - \frac{t^z \cdot (z+r)}{(1+t)^{z+r+1}} \right] \\
&= \sum_{z=k}^{\infty} \left[\frac{(z+r-1)!}{z! \cdot (r-1)!} \cdot \frac{z \cdot t^{z-1}}{(1+t)^{z+r}} - \frac{(z+r-1)!}{z! \cdot (r-1)!} \cdot \frac{(z+r) \cdot t^z}{(1+t)^{z+r+1}} \right] \\
&= \sum_{z=k}^{\infty} \left[\frac{(z+r-1)!}{(z-1)! \cdot (r-1)!} \cdot \frac{t^{z-1}}{(1+t)^{z+r}} - \frac{(z+r)!}{z! \cdot (r-1)!} \cdot \frac{t^z}{(1+t)^{z+r+1}} \right] \\
&= \frac{(k+r-1)!}{(k-1)! \cdot (r-1)!} \cdot \frac{t^{k-1}}{(1+t)^{k+r}} - \\
&\quad - \frac{(k+r)!}{k! \cdot (r-1)!} \cdot \frac{t^k}{(1+t)^{k+r+1}} + \frac{(k+r)!}{k! \cdot (r-1)!} \cdot \frac{t^k}{(1+t)^{k+r+1}} - \\
&\quad - \frac{(k+r+1)!}{(k+1)! \cdot (r-1)!} \cdot \frac{t^{k+1}}{(1+t)^{k+r+2}} + \frac{(k+r+1)!}{(k+1)! \cdot (r-1)!} \cdot \frac{t^{k+1}}{(1+t)^{k+r+2}} + \cdots \\
&= \frac{(k+r-1)!}{(k-1)! \cdot (r-1)!} \cdot \frac{t^{k-1}}{(1+t)^{k+r}}
\end{aligned}$$

where the last equation follows since all the intermediate terms cancel out except the first and the last term. However, the last term converges to 0 as $k \to \infty$. Hence,

$$\begin{aligned} -\frac{\partial F_Z(\lfloor z \rfloor)}{\partial t} &= \frac{\Gamma(k+r)}{\Gamma(k)\cdot\Gamma(r)} \cdot \frac{t^{k-1}}{(1+t)^{k+r}} \\ &= \frac{1}{B(r,k)} \cdot \frac{t^{k-1}}{(1+t)^{k+4}} \end{aligned} \tag{2.35}$$

Then, integration yields,

$$\begin{aligned} -F_Z\left(k-1\right) &= \frac{1}{B(r,k)} \cdot \int_0^t x^{k-1} \cdot (1+x)^{-(k+r)} dx + C \\ &= \frac{1}{B(r,k)} \cdot \int_0^{\frac{t}{1+t}} (\frac{p}{1-p})^{k-1} \cdot (\frac{1}{1-p})^{-(k+r)} \cdot \frac{1}{(1-p)^2} dp + C^{'} \\ &= \frac{1}{B(r,k)} \cdot \int_0^{\frac{t}{1+t}} p^{k-1} \cdot (1-p)^{r-1} dp + C^{'} \\ &= \frac{1}{B(r,k)} \cdot B\left(k, r; \frac{t}{1+t}\right) + C^{'}, \text{ by definition of Incomplete Beta function.} \\ \Rightarrow -F_Z(k-1) &= \frac{B\left(k, r; \frac{t}{1+t}\right)}{B\left(r,k\right)} + C^{'}, \text{ set } \frac{t}{1+t} = p. \\ \Rightarrow -F_Z(k-1) &= \frac{B\left(k,r;p\right)}{B\left(r,k\right)} + C^{'} \\ \Rightarrow -\sum_{z=0}^{k-1} \binom{z+r-1}{z} \cdot p^z \cdot (1-p)^r &= \frac{B(k,r;p)}{B(k,r)} + C^{'} \end{aligned}$$

where $B(k,r) = B(r,k)$ by property of the binomial function.

Take $p = 0$, the last equation becomes

$$\begin{aligned}
-1 &= 0 + C^{'} \\
&\Rightarrow C^{'} = -1 \\
&\Rightarrow -F_Z(k-1) = \frac{B(k,r;p)}{B(r,k)} - 1 \\
&\Rightarrow F_Z(k-1) = 1 - \frac{B(k,r;p)}{B(r,k)} \\
&= 1 - I_p(k,r) \\
&= I_{1-p}(r,k) \\
&\Rightarrow F_Z(k-1) = P(Z \leq k-1) \\
&= P(Z \leq \lfloor z \rfloor) \\
&= P(Z \leq z) \\
&= I_{1-p}(r,k) \\
&\Rightarrow F_Z(z) = I_{1-p}(r, \lfloor z \rfloor + 1)
\end{aligned}$$

where $I_x(a,b)$ is the regularized incomplete Beta function with the property: $I_x(a,b) = 1 - I_{1-x}(b,a)$

□

Interpolation for Zero-Inflated Model

Proof. Suppose Y follows a discrete distribution that can only take non-negative integer values. Let $\tilde{Y}$ be the ZI counterpart of Y. Then, For $y \geq 0$

$$
\begin{aligned}
F_{\tilde{Y}}(y) &= P(\tilde{Y} \leq y) \\
&= \sum_{n=0}^{\lfloor y \rfloor} P(Y = n), \text{ since Y can only take integer values} \\
&= P(Y = 0) + \sum_{n=1}^{\lfloor y \rfloor} P(Y = n) \\
&= p_0 + (1 - p_0) \cdot P(Y = 0) + \sum_{n=1}^{\lfloor y \rfloor} (1 - p_0) \cdot P(Y = n) \\
&= p_0 + (1 - p_0) \cdot \sum_{n=0}^{\lfloor y \rfloor} P(Y = n) \\
&= p_0 + (1 - p_0) \cdot P(Y \leq \lfloor y \rfloor) \\
&= p_0 + (1 - p_0) \cdot F_Y(\lfloor y \rfloor) \\
&= p_0 + (1 - p_0) \cdot F_Y(y)
\end{aligned}
$$

□

Validity of CDF for ZI Model

Proof. For $\pi_{it} = 0$, $G_{y'_{it}}(y) = k(y, \theta_{it})$ where $k(y, \theta_{it}) = F_{y'_{it}}(y)$ is a valid cumulative distribution function (Ilienko, 2013b; Padellini and Rue, 2019a).

For $\pi_{it} > 0$, $G_{y'_{it}}(y) = \pi_{it} + (1 - \pi_{it})k(y, \theta_{it})$. Since $k(y, \theta_{it})$ is a valid cumulative distribution function, functions (3.3) and (3.4) satisfy the following:

$$\lim_{y \to -\infty} G_{y'_{it}}(y) = \lim_{y \to -\infty} [\pi_{it} + (1 - \pi_{it}) \cdot k(y, \theta_{it})] \cdot I_{\{y \geq 0\}} = 0$$

and

$$
\begin{aligned}
\lim_{y \to \infty} G_{y'_{it}}(y) &= \lim_{y \to \infty} [\pi_{it} + (1 - \pi_{it}) \cdot k(y, \theta_{it})] \cdot I_{\{y \geq 0\}} \\
&= \pi_{it} + (1 - \pi_{it}) \cdot \lim_{y \to \infty} k(y, \theta_{it}) \\
&= \pi_{it} + (1 - \pi_{it}) = 1.
\end{aligned}
$$

Moreover, $G_{y'_{it}}(y)$ is non-decreasing since $k(y, \theta_{it})$ is non-decreasing and $(1 - \pi_{it}) \geq 0$, and $G_{y'_{it}}(y)$ is right-continuous since $k(y, \theta_{it})$ is right-continuous. □

Identifiability

Proof. Following Li (2012), to show the identifiability for the continuous counterpart of ZIP model, it is sufficient to show that,

$$f(y; p_1(x), \lambda_1(x), x) = f(y; p_2(x), \lambda_2(x), x) \Rightarrow \lambda_1(x) = \lambda_2(x) \text{ and } p_1(x) = p_2(x) \tag{2.36}$$

For simplicity, denote $p_1(x), p_2(x), \lambda_1(x), \lambda_2(x)$ by $p_1, p_2, \lambda_1, \lambda_2$, respectively. Following Ilienko (2013a) , the density of the continuous Poisson distribution is given by the form,

$$f(y) = c_\lambda \frac{e^{-\lambda}\lambda^y}{\Gamma(y+1)}, \text{ where y} \geq 0 \tag{2.37}$$

where c_λ is a normalizing constant.

Hence, the density for the continuous counterpart of ZIP model is given by the form,

$$f(y) = (1-p)\cdot I_{\{y=0\}} + p \cdot c_\lambda \frac{e^{-\lambda}\lambda^y}{\Gamma(y+1)}, \text{y} \geq 0. \tag{2.38}$$

where p is the probability that the observation Y is from the count process. That is, 1-p is the probability that the observation is from the zero process.

$$\begin{aligned}
&(1-p_1)\cdot I_{\{y=0\}} + p_1 \cdot c_{\lambda_1}\frac{e^{-\lambda_1}\lambda_1^y}{\Gamma(y+1)} = (1-p_2)\cdot I_{\{y=0\}} + p_2 \cdot c_{\lambda_2}\frac{e^{-\lambda_2}\lambda_2^y}{\Gamma(y+1)} \\
&\Rightarrow p_1 \cdot (I_{\{y=0\}} - c_{\lambda_1}\frac{e^{-\lambda_1}\lambda_1^y}{\Gamma(y+1)}) = p_2 \cdot (I_{\{y=0\}} - c_{\lambda_2}\frac{e^{-\lambda_2}\lambda_2^y}{\Gamma(y+1)}) \\
&\Rightarrow \frac{p_1}{p_2} = \frac{I_{\{y=0\}} - c_{\lambda_2}\frac{e^{-\lambda_2}\lambda_2^y}{\Gamma(y+1)}}{I_{\{y=0\}} - c_{\lambda_1}\frac{e^{-\lambda_1}\lambda_1^y}{\Gamma(y+1)}}, \text{ the LHS is a function of } x. \\
&\Rightarrow c(x) = \frac{I_{\{y=0\}} - c_{\lambda_2}\frac{e^{-\lambda_2}\lambda_2^y}{\Gamma(y+1)}}{I_{\{y=0\}} - c_{\lambda_1}\frac{e^{-\lambda_1}\lambda_1^y}{\Gamma(y+1)}} \\
&\Rightarrow I_{\{y=0\}} - c_{\lambda_2}\cdot\frac{e^{-\lambda_2}\cdot\lambda_2^y}{\Gamma(y+1)} = c(x)\cdot I_{\{y=0\}} - c(x)\cdot c_{\lambda_1}\cdot\frac{e^{-\lambda_1}\cdot\lambda_1^y}{\Gamma(y+1)} \\
&\Rightarrow c_{\lambda_2}\cdot e^{-\lambda_2}\cdot\lambda_2^y = [1-c(x)]\cdot I_{\{y=0\}}\cdot\Gamma(y+1) + c(x)\cdot c_{\lambda_1}\cdot e^{-\lambda_1}\cdot\lambda_1^y
\end{aligned}$$

This equation holds for any $y \geq 0$, so,

$$\begin{cases} c_{\lambda_2} \cdot e^{-\lambda_2} \cdot \lambda_2 = c(x) \cdot c_{\lambda_1} \cdot e^{-\lambda_1} \cdot \lambda_1, & \text{if } y = 1. \\ c_{\lambda_2} \cdot e^{-\lambda_2} \cdot \lambda_2^2 = c(x) \cdot c_{\lambda_1} \cdot e^{-\lambda_1} \cdot \lambda_1^2, & \text{if } y = 2. \end{cases} \tag{2.39}$$

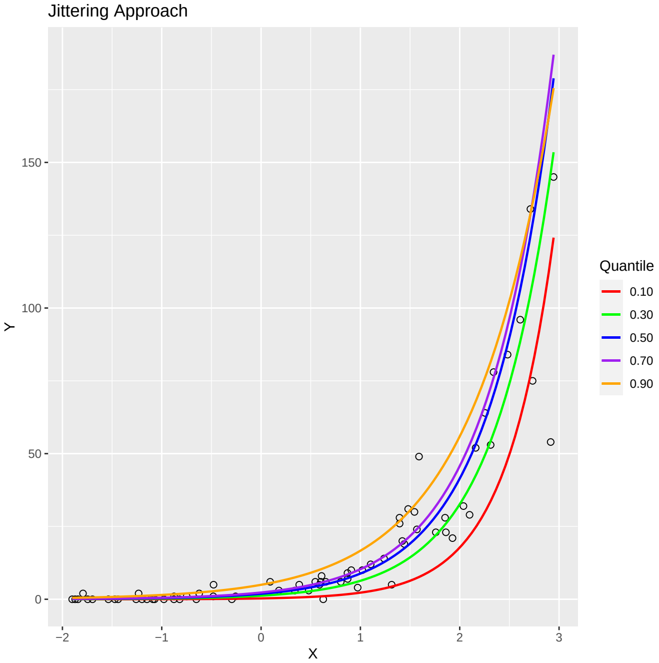
Insert the first equation into the second equation yields,

$$\begin{aligned} & c_{\lambda_2} \cdot e^{-\lambda_2} \cdot \lambda_2 \cdot \lambda_2 = c(x) \cdot c_{\lambda_1} \cdot e^{-\lambda_1} \cdot \lambda_1^2 \\ & \Rightarrow c(x) \cdot c_{\lambda_1} \cdot e^{-\lambda_1} \cdot \lambda_1 \cdot \lambda_2 = c(x) \cdot c_{\lambda_1} \cdot e^{-\lambda_1} \cdot \lambda_1^2 \\ & \Rightarrow \lambda_2 = \lambda_1 \end{aligned}$$

Thus, can see that $\lambda_1(x) = \lambda_2(x)$ and $c(x) = 1$. Since $c(x) = p_1(x)/p_2(x) = 1$, $p_1(x) = p_2(x)$. Hence, the identifiability for the continuous counterpart of ZIP.

Similar extensions to other cases as in Li (2012) can also be obtained. □

Additional Figures



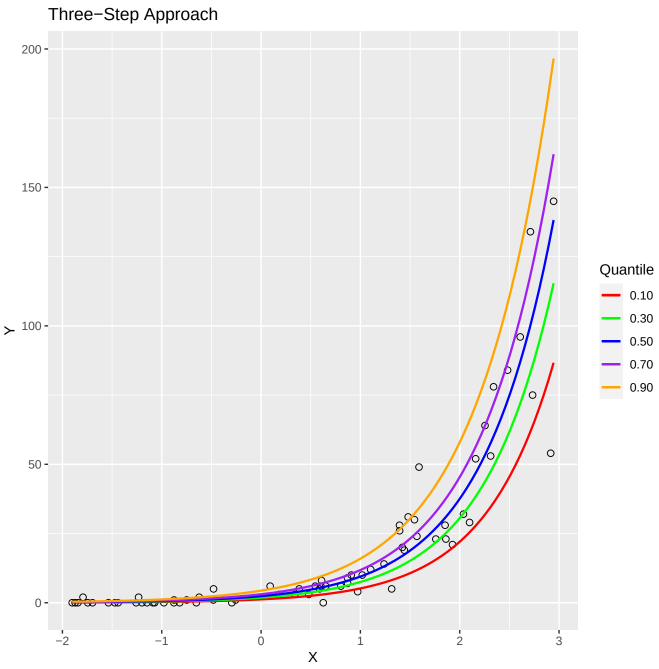


Figure 2.10: Comparison of three-step method and jittering-based method with respect to quantiles crossings.

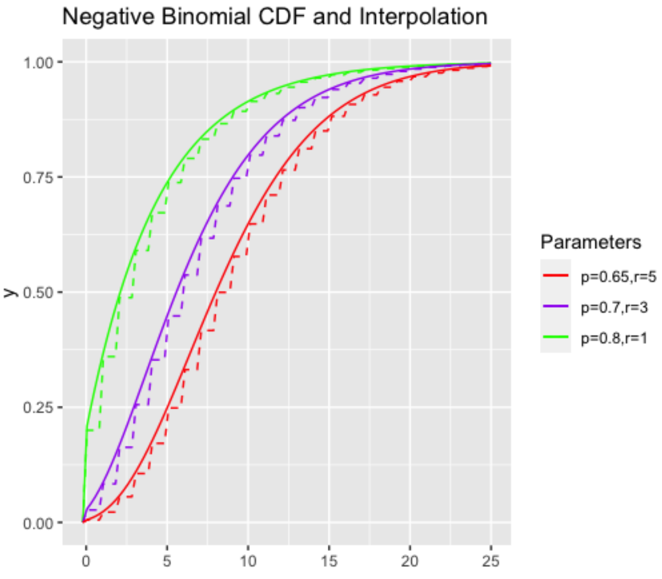


Figure 2.11: Negative Binomial CDF(dash line) and continuous Negative Binomial CDF(solid curve).

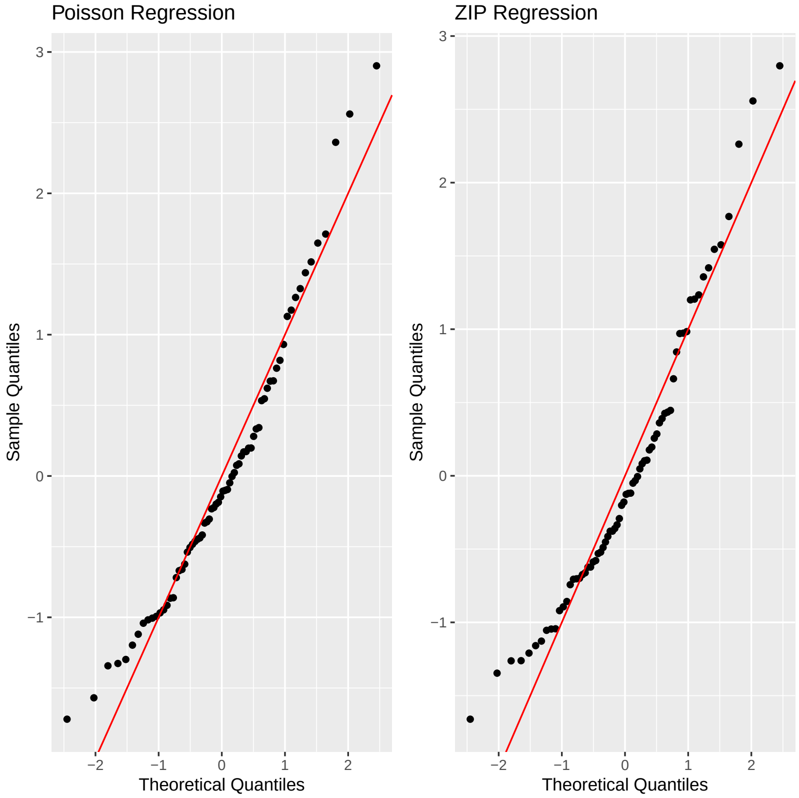


Figure 2.12: Randomized quantile residuals when data generated from Poisson distribution. The left plot is for Poisson regression model (correct model); the right plot is for ZIP regression model (incorrect model).

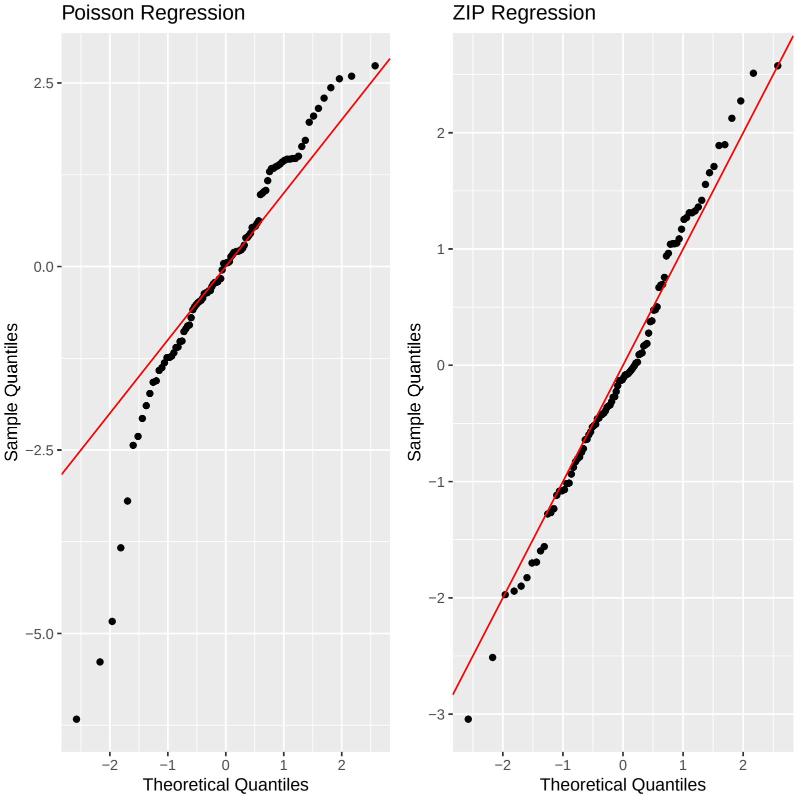


Figure 2.13: Randomized quantile residuals when data generated from ZIP distribution with $p(0) = 0.10$. The left plot is for Poisson regression model (incorrect model); the right plot is for ZIP regression model (correct model).

Simulation Study: Model Misspecification

As can be seen from the description, the three-step approach is fully parametric and requires the specification of a discrete distribution. In order to explore the issue of model mis-specification, simulations were performed with both correct and incorrect distributions. In the following two tables, data were generated from Poisson or ZIP distribution. three-step approaches with Poisson (correct model) and Negative Binomial (model misspecification) are fitted to the data as before. A smaller value is MISE indicates better fit to the data.

In general, when the data is generated from Poisson distribution, models fitted assuming negative binomial distribution does not show huge drop in the performance. This is not very surprising since the negative binomial distribution has more flexibility than the Poisson. On the other hand, when the data is generated from negative binomial distribution, model fitted by Poisson is less robust.

Table 2.25: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as Poisson distribution. Results based on a sample size $n = 250$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	MISE
0.10	TS-NLS(Poisson)	-0.280	0.642	0.206
	TS-NLS(Negative Binomial)	-0.284	0.641	0.222
	JB	-0.244	0.645	0.489
0.25	TS-NLS(Poisson)	0.196	0.574	0.180
	TS-NLS(Negative Binomial)	0.194	0.573	0.203
	JB	0.238	0.570	0.312
0.50	TS-NLS(Poisson)	0.649	0.510	0.176
	TS-NLS(Negative Binomial)	0.648	0.510	0.178
	JB	0.638	0.513	0.230
0.75	TS-NLS(Poisson)	1.016	0.461	0.196
	TS-NLS(Negative Binomial)	1.016	0.461	0.201
	JB	0.972	0.468	0.323
0.90	TS-NLS(Poisson)	1.295	0.426	0.230
	TS-NLS(Negative Binomial)	1.297	0.427	0.543
	JB	1.243	0.433	0.529

Table 2.26: Comparison of point estimates and MISE values obtained by three methods when the response is distributed as Poisson distribution contaminated by 10% of zero-inflation. Results based on a sample size $n = 250$.

τ	Method	$b_0(\tau)$	$b_1(\tau)$	MISE
0.10	TS-NLS(Poisson)	-0.281	0.642	0.214
	TS-NLS(Negative Binomial)	-0.286	0.641	0.233
	JB	-0.923	0.455	25.319
0.25	TS-NLS(Poisson)	0.196	0.573	0.189
	TS-NLS(Negative Binomial)	0.194	0.573	0.201
	JB	-0.038	0.615	0.770
0.50	TS-NLS(Poisson)	0.647	0.510	0.188
	TS-NLS(Negative Binomial)	0.647	0.510	0.194
	JB	0.540	0.528	0.426
0.75	TS-NLS(Poisson)	1.017	0.460	0.215
	TS-NLS(Negative Binomial)	1.018	0.461	0.453
	JB	0.922	0.475	0.562
0.90	TS-NLS(Poisson)	1.295	0.425	0.247
	TS-NLS(Negative Binomial)	1.296	0.427	0.351
	JB	1.214	0.436	0.755

Table 2.27: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as Poisson distribution contaminated by 45% of zero-inflation. Results based on a sample size $n = 1000$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.648(0.045)	0.509(0.012)	0.126
	JB	-0.698(0.292)	0.730(0.099)	16.196
0.75	TS-NLS	1.015(0.040)	0.461(0.011)	0.136
	JB	0.618(0.066)	0.521(0.018)	4.045
0.90	TS-NLS	1.293(0.036)	0.426(0.010)	0.157
	JB	1.054(0.052)	0.459(0.015)	2.821

Table 2.28: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as Poisson distribution contaminated by 45% of zero-inflation. Results based on a sample size $n = 2500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.647(0.027)	0.510(0.007)	0.098
	JB	-0.696(0.202)	0.737(0.050)	14.680
0.75	TS-NLS	1.017(0.026)	0.460(0.007)	0.104
	JB	0.621(0.041)	0.520(0.011)	4.026
0.90	TS-NLS	1.297(0.025)	0.425(0.007)	0.124
	JB	1.059(0.035)	0.458(0.010)	2.705

Table 2.29: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as Poisson distribution contaminated by 10% of zero-inflation. Results based on a sample size $n = 1000$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.648(0.036)	0.509(0.010)	0.112
	JB	0.543(0.047)	0.527(0.013)	0.298
0.75	TS-NLS	1.016(0.030)	0.461(0.008)	0.115
	JB	0.924(0.039)	0.474(0.010)	0.371
0.90	TS-NLS	1.296(0.028)	0.425(0.008)	0.136
	JB	1.219(0.043)	0.435(0.012)	0.507

Table 2.30: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as Poisson distribution contaminated by 45% of zero-inflation. Results based on a sample size $n = 2500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.649(0.021)	0.510(0.006)	0.093
	JB	0.541(0.029)	0.527(0.008)	0.264
0.75	TS-NLS	1.016(0.020)	0.461(0.005)	0.097
	JB	0.922(0.026)	0.475(0.007)	0.345
0.90	TS-NLS	1.296(0.018)	0.425(0.005)	0.114
	JB	1.218(0.027)	0.436(0.008)	0.439

Table 2.31: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as Poisson distribution. Results based on a sample size $n = 1000$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.650(0.031)	0.509(0.009)	0.105
	JB	0.640(0.041)	0.512(0.011)	0.121
0.75	TS-NLS	1.016(0.029)	0.460(0.008)	0.111
	JB	0.971(0.037)	0.468(0.010)	0.178
0.90	TS-NLS	1.294(0.024)	0.426(0.007)	0.129
	JB	1.247(0.037)	0.432(0.011)	0.305

Table 2.32: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as Poisson distribution. Results based on a sample size $n = 2500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.646(0.020)	0.510(0.005)	0.093
	JB	0.637(0.026)	0.513(0.007)	0.101
0.75	TS-NLS	1.016(0.017)	0.461(0.005)	0.095
	JB	0.972(0.022)	0.468(0.006)	0.153
0.90	TS-NLS	1.295(0.016)	0.425(0.004)	0.111
	JB	1.247(0.025)	0.432(0.007)	0.260

Table 2.33: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as negative binomial distribution contaminated by 45% of zero-inflation. Results based on a sample size $n = 1000$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.468(0.091)	0.511(0.026)	0.351
	JB	-1.229(0.232)	0.529(0.168)	50.828
0.75	TS-NLS	1.051(0.082)	0.491(0.026)	0.743
	JB	0.449(0.127)	0.534(0.041)	27.822
0.90	TS-NLS	1.487(0.079)	0.477(0.026)	1.564
	JB	1.173(0.092)	0.494(0.030)	25.301

Table 2.34: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as negative binomial distribution contaminated by 45% of zero-inflation. Results based on a sample size $n = 2500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.476(0.059)	0.508(0.017)	0.200
	JB	-1.280(0.140)	0.567(0.097)	48.972
0.75	TS-NLS	1.057(0.050)	0.489(0.017)	0.357
	JB	0.459(0.078)	0.532(0.025)	27.674
0.90	TS-NLS	1.488(0.053)	0.478(0.017)	0.722
	JB	1.169(0.065)	0.495(0.022)	24.530

Table 2.35: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as negative binomial distribution contaminated by 10% of zero-inflation. Results based on a sample size $n = 1000$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.470(0.077)	0.509(0.022)	0.252
	JB	0.292(0.093)	0.525(0.030)	1.290
0.75	TS-NLS	1.060(0.064)	0.488(0.020)	0.444
	JB	0.940(0.072)	0.501(0.024)	1.630
0.90	TS-NLS	1.493(0.063)	0.477(0.021)	0.956
	JB	1.414(0.076)	0.483(0.024)	2.454

Table 2.36: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as negative binomial distribution contaminated by 10% of zero-inflation. Results based on a sample size $n = 2500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.469(0.044)	0.510(0.013)	0.149
	JB	0.288(0.060)	0.527(0.019)	1.125
0.75	TS-NLS	1.060(0.039)	0.489(0.012)	0.233
	JB	0.938(0.046)	0.501(0.015)	1.413
0.90	TS-NLS	1.488(0.041)	0.478(0.013)	0.456
	JB	1.407(0.047)	0.485(0.015)	1.957

Table 2.37: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as negative binomial distribution. Results based on a sample size $n = 1000$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.466(0.061)	0.511(0.019)	0.225
	JB	0.449(0.078)	0.516(0.025)	0.315
0.75	TS-NLS	1.060(0.056)	0.489(0.019)	0.402
	JB	1.011(0.0068)	0.498(0.022)	0.592
0.90	TS-NLS	1.486(0.063)	0.479(0.021)	0.869
	JB	1.448(0.072)	0.485(0.024)	1.469

Table 2.38: Comparison of point estimates and MISE values obtained by two methods when the response is distributed as negative binomial distribution. Results based on a sample size $n = 2500$.

τ	Method	$b_0(\tau)(SE)$	$b_1(\tau)(SE)$	MISE
0.50	TS-NLS	0.468(0.038)	0.510(0.012)	0.138
	JB	0.448(0.050)	0.516(0.016)	0.181
0.75	TS-NLS	1.060(0.034)	0.489(0.011)	0.221
	JB	1.017(0.041)	0.496(0.014)	0.341
0.90	TS-NLS	1.488(0.034)	0.478(0.011)	0.408
	JB	1.451(0.046)	0.483(0.015)	0.750

Simulation Study: Regression Setting

The following Monte Carlo simulation results investigate the performance of our method compared with existing method in a regression setting. The comparison is similar to the previous unconditional scenario. However, the unconditional scenario deals with only one distribution (for example, $Poisson(3)$ in previous simulation); on the other hand, the regression setting deals with n conditional distributions. For each unique level of the predictors, there is a distribution with different values of parameter.

Another distinction to make is that, in the main content of the paper, a parametric bootstrap is employed to estimate the SE of the regression coefficients. Hence, the focus of the inference is on the $\boldsymbol{\beta}(\tau)$. In the following simulation setting, the goal is to construct bootstrap confidence intervals for the response values and the inference is on the $\mathbf{Y}$.

The first measurement is the Mean Squared Error of Prediction. This is defined as the MSE between the true conditional quantile, $Q_\tau(Y|X = x)$ and the predicted values, $\hat{Q}_\tau(Y|X = x)$.

$$\frac{1}{n}\sum_{i=1}^{n}[Q_\tau(Y|X = x_i) - \hat{Q}_\tau(Y|X = x_i)]^2 \tag{2.40}$$

This gives us a straightforward measurement of how close each method predicts the conditional quantiles.

The other two measurements are the coverage probability and the average length for the bootstrap confidence intervals based on each method. In the main content of the paper, a parametric bootstrap is employed to estimate the SE of the regression coefficients. In the following simulation setting, the goal is to construct bootstrap confidence intervals for the response values. a paired bootstrap will be used instead of the parametric bootstrap to construct the bootstrap confidence intervals. That is, we obtain a bootstrap sample by sampling with replacement from the pairs, $(y_1, x_1), ..., (y_n, x_n)$. In total, each bootstrap confidence interval is calculated by averaging B bootstrap samples. Note that each bootstrap sample could have different values of covariate. Thus, to guarantee the average is meaningful across B samples, we proposed the following routine:

1. In each simulation j (j = 1, ..., N) with a specific sample size n, generate the training covariate, $x_1, ..., x_n$ from a uniform distribution over $(0, 1)$. Then, for a specified proportion of ZI, π_0, generate the corresponding response, $y_1, ..., y_n$. This yields the original dataset with pairs, $(y_1, x_1), ..., (y_n, x_n)$.

2. Within the same simulation, obtain a sequence of equally-spaced test covariate, $x_1^0, ..., x_n^0$, from the same support, $(0, 1)$. Then, obtain the conditional quantiles at each value of $x_1^0, ..., x_n^0$. The resulting sequence, $Q_\tau(Y|x_1^0), ..., Q_\tau(Y|x_n^0)$, is equally-spaced over the support.

3. Conduct a paired bootstrap by sampling with replacement from the pairs, $(y_1, x_1), ..., (y_n, x_n)$. Within each bootstrap sample, calculate a bootstrap confidence interval for the conditional quantile at each unique value of the test covariate, $x_1^0, ..., x_n^0$.

4. Obtain the pointwise coverage probability and average length at each unique value of the test covariate, $x_1^0, ..., x_n^0$. The results are based on B bootstrap samples within the j^{th} simulation.

5. Calculate the coverage probability and average length at each unique value of the test covariate, $x_1^0, ..., x_n^0$. Average the results from the previous step over N to obtain the empirical coverage probability and average interval length at each unique value of the test covariate, $x_1^0, ..., x_n^0$. The final results are plotted to show the overall performance by each method.

This section extends unconditional distributions to conditional distributions, that is, regression setting. The first scenario is the regular Poisson regression with one covariate. Below is a summary of the MSE by different methods. In the following table, data were simulate from Poisson distribution with mean parameter,

$$\lambda = e^{0.7+0.5*X}, X \sim uniform(0, 1) \tag{2.41}$$

Table 2.39: Comparison of MSE values for different implementations. Data generated from Poisson distribution. Results based on 1000 simulations.

N	τ	DR	NLS	⌈NLS⌉	Jit	⌊Jit⌋
100	0.50	0.19	0.35	0.14	0.40	0.18
	0.75	0.27	0.35	0.19	0.52	0.24
	0.90	0.42	0.40	0.24	0.64	0.38
400	0.50	0.08	0.32	0.06	0.35	0.07
	0.75	0.11	0.29	0.08	0.44	0.11
	0.90	0.19	0.34	0.13	0.46	0.18
1600	0.50	0.03	0.31	0.03	0.33	0.03
	0.75	0.05	0.27	0.04	0.42	0.06
	0.90	0.10	0.32	0.07	0.43	0.11

As can be seen from the above table, the three-step approach with NLS routine ($\lceil NLS \rceil$) performed well in estimating the true quantiles. This further confirms the validity of the estimation for the regression coefficients.

When the data exhibits certain degree of ZI (tables on the following pages), the three-step approach provided even greater advantages over competing methods. Hence, when the goal is to infer about the conditional quantiles of the count process, the three-step approach provides a consistent mechanism to distinguish responses from different sources.

Table 2.40: Comparison of MSE values for different implementations. Data generated from ZIP distribution with $\pi_0 = 0.15$. Results based on 1000 simulations.

N	τ	DR	NLS	$\lceil$NLS$\rceil$	Jit	$\lfloor$Jit$\rfloor$
100	0.50	0.39	0.40	0.19	0.19	0.37
	0.75	0.36	0.40	0.26	0.31	0.32
	0.90	0.49	0.47	0.31	0.46	0.43
400	0.50	0.32	0.33	0.07	0.12	0.31
	0.75	0.20	0.30	0.10	0.22	0.17
	0.90	0.23	0.35	0.15	0.29	0.20
1600	0.50	0.30	0.31	0.03	0.10	0.29
	0.75	0.16	0.27	0.04	0.20	0.13
	0.90	0.19	0.33	0.08	0.25	0.15

Table 2.41: Comparison of MSE values for different implementations. Data generated from ZIP distribution with $\pi_0 = 0.45$. Results based on 1000 simulations.

N	τ	DR	NLS	$\lceil$NLS$\rceil$	Jit	$\lfloor$Jit$\rfloor$
100	0.50	3.61	0.44	0.26	1.72	3.52
	0.75	1.33	0.45	0.33	0.46	1.20
	0.90	1.06	0.55	0.41	0.41	0.92
400	0.50	3.58	0.34	0.09	1.79	3.44
	0.75	1.04	0.32	0.13	0.29	1.01
	0.90	0.76	0.38	0.19	0.19	0.71
1600	0.50	3.45	0.31	0.04	1.85	3.21
	0.75	0.98	0.28	0.06	0.24	0.96
	0.90	0.72	0.33	0.10	0.13	0.66

Simulation Study: Prediction Intervals for Partial Effects by Different Bootstrap Implementations

The following tables report the overall coverage probability for the difference in the conditional quantiles. This difference is considered as a partial effect in the econometric field:

$$Q_y(\tau|x_i) - Q_y(\tau|x_0), \tag{2.42}$$

where in this simulation setting, $x_0 = 2.0$ and $x_i = 2.1, 2.2, \cdots, 2.9, 3$.

For example, when data were generated from Poisson distribution without ZI, the 50^{th}conditional quantile at $x_0 = 2$ and $x_i = 2.1$ are 5 and 6, respectively. Hence, the difference in the conditional quantile is 1.

As can be seen from Table 2.42, the overall coverage probability is usually higher than the nominal level, 0.95. One main reason for this performance can be due to the discrete nature of both the response variables and the constructed CIs. To illustrate the difference, similar tables reporting the overall coverage probability in the continuous case are reported.

In Table 2.43, the true quantiles and the difference in conditional quantiles are defined for the underlying continuous Poisson distribution. The bootstrap CIs are constructed by the percentile method. One difference between the CIs for the discrete case and the CIs for the continuous case is that, the ceiling transformation $\lceil x \rceil$ is applied to the predictions for the discrete case, while the predictions for the continuous case are used directly.

Table 2.42: Coverage probability based on different implementations of pairwise bootstrap and multiplier(weighted) bootstrap; data generated from Poisson distribution.

π_0	τ	$\lceil$Pairwise$\rceil$	$\lceil$Multiplier$\rceil$
0	0.50	0.990	0.992
	0.75	0.997	0.996
	0.90	0.978	0.983
0.15	0.50	0.994	0.994
	0.75	0.995	0.994
	0.90	0.980	0.974
0.45	0.50	0.989	0.988
	0.75	0.998	0.997
	0.90	0.984	0.982

Table 2.43: Coverage probability based on different implementations of pairwise bootstrap and multiplier(weighted) bootstrap; data generated from continuous Poisson distribution.

π_0	τ	Pairwise	Multiplier
0	0.50	0.933	0.937
	0.75	0.940	0.942
	0.90	0.920	0.930
0.15	0.50	0.940	0.932
	0.75	0.940	0.930
	0.90	0.930	0.940
0.45	0.50	0.937	0.940
	0.75	0.950	0.940
	0.90	0.940	0.940

The following tables report the overall coverage probability for the difference in the conditional quantiles where $x_0 = 5.0$ and $x_i = 5.1, 5.2, \cdots, 5.9, 6$.

For example, when data were generated from Poisson distribution without ZI, the 50^{th}conditional quantile at $x_0 = 5$ and $x_i = 5.1$ are 24 and 26, respectively. Hence, the difference in the conditional quantile is 2.

Table 2.44: Coverage probability based on different implementations of pairwise bootstrap and multiplier(weighted) bootstrap; data generated from Poisson distribution.

π_0	τ	$\lceil$Pairwise$\rceil$	$\lceil$Multiplier$\rceil$
0	0.50	0.990	0.987
	0.75	0.993	0.994
	0.90	0.988	0.987

Table 2.45: Coverage probability based on different implementations of pairwise bootstrap and multiplier(weighted) bootstrap; data generated from continuous Poisson distribution.

π_0	τ	Pairwise	Multiplier
0	0.50	0.940	0.949
	0.75	0.940	0.940
	0.90	0.940	0.933

The following tables report the overall coverage probability for the difference in the conditional quantiles with data generated from negative binomial/ZINB distribution.

Table 2.46: Coverage probability based on different implementations of pairwise bootstrap and multiplier(weighted) bootstrap; data generated from negative binomial distribution with different proportions of zero-inflation.

π_0	τ	⌈Pairwise⌉	⌈Multiplier⌉
0	0.50	0.999	0.999
	0.75	0.996	0.997
	0.90	0.982	0.983
0.15	0.50	0.994	0.994
	0.75	0.995	0.994
	0.90	0.980	0.974
0.45	0.50	0.989	0.988
	0.75	0.998	0.997
	0.90	0.984	0.982

Chapter 3 Quantile Functions for Zero-Inflated Longitudinal Count Data

3.1 Introduction

As introduced in the previous chapter, QR is a widely used approach to estimate flexible models in economics and statistics. While theoretical and methodological research in the last 40 years has been addressing essential generalizations of the original approach (Koenker, 2017), the literature on the analysis of discrete data remains open to challenges and possibilities. In many applications, practitioners face the limitations of classical parametric models, where the effect of a treatment variable can be heterogeneous throughout the conditional distribution of the count variable. However, policy recommendations can only be based on average effects. See Cameron and Trivedi (2013) for a detailed summary of econometric analysis with count data.

An illustrative example includes the number of visits to physicians and the demand for medical services. Using RAND Health Insurance Experiment data (Deb and Trivedi, 2002), Figure 3.1 shows that the proportion of zero visits to physicians exceeds 30% for patients with no greater than 15 visits per year. Moreover, the count response distribution has a long tail reaching a maximum of 77 visits, while the average is 2.86. As in many other applications, the need for a flexible approach that simultaneously addresses zero inflation and latent subject heterogeneity while allowing estimation of effects across the conditional distribution is immediately apparent.

This chapter investigates the estimation of conditional quantile functions and covariate effects for discrete responses in a longitudinal setting. Our approach is based on a continuous approximation to distribution functions for count data within a class of models commonly employed in the literature. We adopt an approach based on interpolation of functions for discrete responses as in Ilienko (2013b) and Padellini and Rue (2019a). This provides an alternative smoothing method to the jittering approach proposed by Machado and Santos Silva (2005) and adopted by Harding and Lamarche (2019a). We extend the three-step estimation procedure in Chapter 2, which provides a flexible statistical frame-

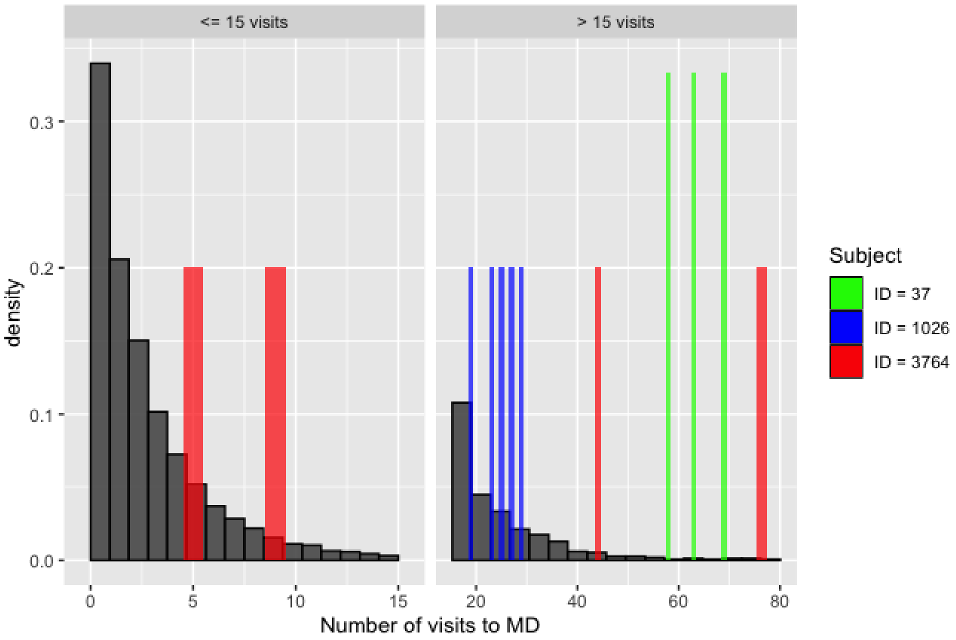


Figure 3.1: Number of visits to physicians in the RAND Health Insurance Experiment data.

work to handle over/underdispersion, shrinkage estimation and smoothing of regression relationships. In the first step, we consider the estimation of the conditional mean model. In the second step, we obtain a conditional quantile variate as the solution of a nonlinear moment condition defined for the conditional mean. We show that the solution exists and it is unique. Finally, in a third step, interpolation is employed to model conditional quantile responses. The estimator's finite sample performance is investigated using a simulation study, and we find that the estimator has a satisfactory performance for the estimation of quantile effects under different degrees of zero inflation.

Our work is based on the previous chapter and is related to the recent research that has contributed to the generalization of conditional quantile models for count data. The original work of Machado and Santos Silva (2005) introduced a jittering approach to smooth the count response variable. Lee and Neocleous (2010) proposed a Bayesian approach, and Chernozhukov, Fernández-Val, and Weidner (2017) develop an approach based on distribution regression. The literature on panel quantiles includes just a few papers. Harding and Lamarche (2019a) extend the jittering approach to longitudinal data without zero inflation, and Wang, Wu, Zhao, and Zhou (2020) propose an estimator for time-varying coefficients using a quadratic inference function approach within a quantile framework. The estimator proposed in this chapter is different from existing approaches for two important reasons. First, existing QR approaches have not been developed for ZI models with longitudinal data. Second, we consider estimating the conditional mean model in the first step, rather than considering a QR model as in Padellini and Rue (2019a). Therefore, the proposed methodology allows flexibility to estimate a class of models with subject heterogeneity, without considerations on the minimum number of repeated observations per subject as in panel data QR models (Harding and Lamarche, 2019a).

As highlighted in the preceding chapter, one of this work's contributions is addressing zero inflation in longitudinal data. Zero inflation occurs when zero counts arise from one of two possible states: a degenerate state or some discrete probability distribution. This structure is easily modeled using a two-component mixture model. The seminal work by Lambert (1992) is the earliest paper to thoroughly develop the ZIP regression model as a way to characterize zero defects in a manufacturing process as manifesting from one of two states: a *perfect state* and an *imperfect state*. Since then, numerous extensions to the ZIP regression model have been developed; see Young et al. (2021) and Young et al. (2021) for a contemporary review. In particular, just like in non-ZI models, the random effects in ZI models have been used to capture various features of the data, such as subject heterogeneity (Zhu, 2012), serial dependency between successive responses (Yau et al., 2004), and spatial association (Agarwal et al., 2002). Our work is consistent with the spirit of such contributions in that we will be using random individual intercepts accounting for subject heterogeneity when estimating conditional quantiles for longitudinal data with ZI discrete responses.

This chapter is organized as follows. In Section 3.2, we formalize the development of quantiles for ZI count regression models by transferring the problem to one that utilizes the continuous version of the discrete model under consideration. In Section 3.3, we provide details of GLMMs with an emphasis on panel count outcomes, how to incorporate zero inflation, and pose the problem of performing quantile regression in such ZI GLMMs. In Section 3.4, we provide an extensive simulation study to assess our approach's performance in estimating mean and quantile effects. In Section 3.5, we analyze data from the RAND Health Insurance Experiment and provide new insights using our modeling paradigm. We end with a discussion in Section 3.6.

3.2 Conditional Quantiles of Count Responses

In a longitudinal setting, we observe a random sample of N subjects. The i^{th} subject has T_i measured count outcomes, which are collectively represented by the T_i-dimensional vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})^\top$, $i = 1, \ldots, N$. Note that the T_i need not be the same for all units, however, the setting where $T_i \equiv T$ is considered balanced. Henceforth, we only consider the balanced setting to keep notation simple, but everything discussed extends to the unbalanced setting. Associated with the t^{th} record from the i^{th} subject is a vector of p independent variables, given by $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itp})^\top$. Assume further that with the t^{th} record from the i^{th} unit is a vector of q predictor variables (random effects), given by $\mathbf{z}_{it} = (z_{it1}, \ldots, z_{itp})^\top$. It is clear from the above description that all the results in Chapter 2 can be extended to a longitudinal setting. The only difference is the subscript i replaced by it, reflecting the fact that each unit now has t repeated measurements.

Let $\theta_{it} = \mathrm{E}[y_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}]$ denote the conditional mean of the distribution $F_{y_{it}}$ of the count response, y_{it}. Let $G_{y_{it}}$ be the cumulative distribution function of a ZI count variable,

$$G_{y_{it}}(y) = \pi_{it} + (1 - \pi_{it})F_{y_{it}}(y), \tag{3.1}$$

for $y \in \mathbb{N}$, where π_{it} is the probability that the outcome variable has a degenerate distribution at zero. This is the source of extra zeros and the probability π_{it} can be influenced by covariates, as shown below.

As before, we propose to consider the following continuous counterpart to the ZI count distribution (3.1):

$$G_{y'_{it}}(y) = \pi_{it} + (1 - \pi_{it})k(y, \theta_{it}), \tag{3.2}$$

where $k(y, \theta_{it}) = F_{y'_{it}}(y)$ is the CDF of y'_{it}, which is defined as the continuous version of y_{it}. The function $k(\cdot, \theta_{it})$ is continuous and increasing in its first argument, and it satisfies

$k(\lfloor y \rfloor, \theta_{it}) = F_{y_{it}}(y)$, where the floor function $\lfloor x \rfloor := \max\{y \in Z : y \leq x\}$. See Chapter 2 for similar derivations in the setting without repeated measures.

The extension (3.2) can be used on the two leading distributions: the ZIP and ZINB regression model, in the same way as in Chapter 2. If $y_{it} \sim ZIP(\theta, \pi)$, then

$$G_{y_{it}}(y) = \pi_{it} + (1 - \pi_{it}) \frac{\Gamma(\lfloor y \rfloor + 1, \theta_{it})}{\Gamma(\lfloor y \rfloor + 1)},$$

where $\Gamma(x, \theta) = \int_\theta^\infty e^{-s} s^{x-1} ds$ denotes the upper incomplete gamma function. It follows then, for $y > -1$,

$$G_{y'_{it}}(y) = \pi_{it} + (1 - \pi_{it}) \frac{\Gamma(y + 1, \theta_{it})}{\Gamma(y + 1)}. \tag{3.3}$$

On the other hand, if $y_{it} \sim ZINB(r, p_{it})$, where r is the number of failures in a series of Bernoulli trials and $p_{it} \in (0, 1)$ is the probability of success, we have

$$G_{y'_{it}}(y) = \pi_{it} + (1 - \pi_{it}) I_{1-p_{it}}(r, y + 1) = \pi_{it} + (1 - \pi_{it}) \frac{B(r, y + 1, 1 - p_{it})}{B(r, y + 1)}, \tag{3.4}$$

where $I_{1-p_{it}}(r, y + 1)$ is the regularized incomplete beta function and $B(r, y + 1) = \int_0^1 s^r (1 - s)^{-y} ds$ is the beta function.

A logistic regression model is used to model the zero-inflation probability,

$$\begin{aligned} \pi_{it} &= \frac{\exp(\mathbf{w}_{it}^\top \boldsymbol{\gamma})}{(1 + \exp(\mathbf{w}_{it}^\top \boldsymbol{\gamma}))} \\ \theta_{it} &= (1 - \pi_{it}) \exp(\mathbf{x}_{it}^\top \boldsymbol{\beta} + \mathbf{z}_{it}^\top \mathbf{u}_i) = \frac{\exp(\mathbf{x}_{it}^\top \boldsymbol{\beta} + \mathbf{z}_{it}^\top \mathbf{u}_i)}{(1 + \exp(\mathbf{w}_{it}^\top \boldsymbol{\gamma}))}, \\ \tau &= \pi_{it} + (1 - \pi_{it}) \frac{\Gamma(q_{it} + 1, \theta_{it})}{\Gamma(q_{it} + 1)}, \end{aligned}$$

where $\mathbf{w}_{it}$ is a vector of independent variables that can be the same as $\mathbf{x}_{it}$, and $\tau \in (0, 1)$ is the quantile level. The unknowns are the parameters $(\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top, \mathbf{u}_i^\top)^\top$, and the τ-th quantile of the continuous response variable, q_{it}, in a model with mean θ_{it} and probability π_{it}.

In the third step, we obtain the quantile-specific effects on the count response variable from the following problem:

$$\mathrm{E}\{L(q_{it}(\tau) - g(\eta_{it}))\}, \tag{3.5}$$

where $L(\cdot)$ is a loss function, $g(\cdot)$ is a link function, and η_{it} is a predictor variable which is determined by regressors and individual intercepts.

In a longitudinal setting, the computational routine of our three-step model is more complicated than the cross-sectional version in Chapter 2. Due to repeated measures and random effects, the interpolation in the second step requires extra caution. See Lamarche et al. (2021) for relative results for the existence and uniqueness of the solution.

The above procedures offer a general formulation extended to the longitudinal settings. We then provide specific forms for our model in the Section 3.3.

3.3 Model Specification and Estimation

Let the $\mathbf{x}_{it}^{\top}$ be the rows of the $T \times p$ design matrix $\mathbf{X}_i$ and the $\mathbf{z}_{it}^{\top}$ be the rows of the $T \times q$ design matrix $\mathbf{Z}_i$. Both $\mathbf{X}_i$ and $\mathbf{Z}_i$ include a column of 1s for a fixed intercept and a random intercept, respectively. That is, $x_{it1} \equiv 1$ and $z_{it1} \equiv 1$ for all i and t.

In GLMMs, the link function $g(\cdot)$ is used to relate $\mathbf{y}_i$ to the linear predictor

$$\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i, \tag{3.6}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\mathbf{u}_i \in \mathbb{R}^q$ are vectors of fixed-effect coefficients and random-effect coefficients, respectively. Let $\boldsymbol{\mu}_i|\mathbf{u}_i$ denote the mean of the conditional distribution for the response variable. The link function is defined such that $\mathrm{E}[\mathbf{y}_i|\mathbf{u}_i] = (\boldsymbol{\mu}_i|\mathbf{u}_i) = g^{-1}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i|\mathbf{u}_i) = g^{-1}(\boldsymbol{\eta}_i|\mathbf{u}_i)$, where $h(\cdot) \triangleq g^{-1}(\cdot)$ is used to write the inverse link function. Typically, the preceding setup assumes that the $\mathbf{u}_i$ are iid $\mathcal{N}_q(\mathbf{0}, \mathbf{G})$, where $\mathbf{G}$ is positive definite. To further solidify the form of the random component of (3.6) for the application presented in Section 3.5, consider the case where only a random intercept is present. In that setting, $\mathbf{Z}_i \equiv \mathbf{1}_T$ and $\mathbf{u}_i = u_i$, which becomes a scalar. That is, $q = 1$. However, our proposed methodology can be framed using a more general form of $\mathbf{Z}_i$ and $\mathbf{u}_i$, which can be extended to situations with more complex mixed-effects structures.

The GLMM is a type of hierarchical model where the hierarchical structure is characterized through the random effects. Thus, GLMMs allow a natural framework to reflect panels of units, such as repeated measures on the same subject. The framework also accommodates intra-subject correlation, a statistical feature leveraged in longitudinal data setting as in the present chapter.

As in the previous chapter, our work is focused on the setting where the data's response values are counts. Thus, the choices of the Poisson and the negative binomial for $y_{it}|\mathbf{u}_i$ are reasonable. Given the many parameterizations of the negative binomial distribution, it is worth emphasizing again that the Type 2 parameterization is used for the negative binomial (the NB2 model in Hilbe (2011) and Chapter 2).

One challenge with MLE of GLMMs in a longitudinal setting is that the marginal likelihood involves integration over a complicated product of Gaussian and exponential family likelihoods (or quasi-likelihoods). As a result, direct maximization is generally impossible; computationally, integral approximations are done via Gauss-Hermite quadrature or Laplace approximations, both of which typically perform very well. In `R` (R Core Team, 2016), both approximations are options in the `glmer()` function within the `lme4` package (Bates et al., 2015), while only the Laplace approximation is available in the `glmmTMB()` function within the `glmmTMB` package (Brooks et al., 2017). For a thorough treatment of GLMM methodology, we refer to the text by Stroup (2013).

We next describe ZI GLMMs for ZIP and ZINB distributions in the longitudinal setting. For $\mathbf{G}$, the variance-covariance matrix of the random effects $\mathbf{u}_i$, let vech$(\mathbf{G}) \in \Lambda$, where Λ be an open subset of $\mathbb{R}^{q(q+1)/2}$, such that the dimension is determined by the half-vectorization of $\mathbf{G}$. Let $\boldsymbol{\xi} \in \Xi$ generically denote the s-dimensional parameter vector for either the Poisson GLMM or negative binomial GLMM. Specifically, $\boldsymbol{\xi} = \boldsymbol{\beta}$ for the Poisson GLMM and $\boldsymbol{\xi} = (\boldsymbol{\beta}^\top, \phi)^\top$ for the negative binomial GLMM, thus resulting in $s \in \{p, p+1\}$. Here, Ξ is the parameter space, which is an open subset of $\mathbb{R}^s$. Suppose now that the zeroes in our count outcomes are generated from one of two possible processes: a degenerate distribution with probability $\pi_{it} \equiv d(\mathbf{w}_{it}^\top \boldsymbol{\gamma})$ or the count distribution $p_{y_{it}|\mathbf{u}_i}$ with probability $1 - \pi_{it}$. Therefore,

$$y_{it}|\mathbf{u}_i \sim \begin{cases} 0, & \text{with probability } \pi_{it}; \\ p_{y_{it}|\mathbf{u}_i}, & \text{with probability } 1 - \pi_{it}. \end{cases} \tag{3.7}$$

Here, π_{it} is again a probability that determines the source of zero counts from the two states, so $d^{-1}(\cdot)$ is taken as a logit link function as logistic regression model is used. The linearization involves an r-dimensional vector of predictors, $\mathbf{w}_{it}$, and a parameter vector

$\boldsymbol{\gamma} \in \Gamma$, where Γ is an open subset of $\mathbb{R}^r$. Note that $w_{it1} \equiv 1$ and represents the intercept term. The pmf for the ZI count variable defined in (3.7) is thus

$$f_{y_{it}|\mathbf{u}_i}(y_{it}; \mathbf{x}_{it}, \mathbf{w}_{it}, \mathbf{z}_{it}, \mathbf{u}_i, \boldsymbol{\vartheta}) = \begin{cases} \pi_{it} + (1 - \pi_{it}) p_{y_{it}|\mathbf{u}_i}(0; \boldsymbol{\xi}, \mathbf{G}), & \text{if } y_{it} = 0; \\ (1 - \pi_{it}) p_{y_{it}|\mathbf{u}_i}(y_{it}; \boldsymbol{\xi}, \mathbf{G}), & \text{if } y_{it} \in \mathbb{N}^+, \end{cases} \tag{3.8}$$

where $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$ and $\boldsymbol{\vartheta} = (\boldsymbol{\xi}^\top, \boldsymbol{\gamma}^\top, \text{vech}(\mathbf{G})^\top)^\top \in \Theta = \Xi \times \Gamma \times \Lambda$. Maximum likelihood of ZI GLMMs can be done via `TMB` in the `glmmTMB` package, which is how we proceed. In particular, we are able to obtain point estimates of all parameters, which are BLUEs, as well as the best linear unbiased predictors (BLUPs) of the $\mathbf{u}_i$. For a detailed description of ZI GLMMs derivation and methodology, we refer to the text by Lamarche et al. (2021).

Three-Step Estimator

We now summarize the three-step procedure in the longitudinal setting for constructing the estimated quantile effects, $\hat{\boldsymbol{\beta}}^\tau$ in Procedure 3.

Procedure 3 Modeling Procedure for Quantile Count Regression with Zero-Inflated Longitudinal Data

(1) For the assumed ZI GLMM having pmf of the form in (3.8), find the maximum likelihood estimate for the mean parameters $\boldsymbol{\vartheta}$, $\hat{\boldsymbol{\vartheta}}$, using $\mathcal{L}^*(\boldsymbol{\vartheta})$, the Laplace approximation for the marginal likelihood as defined in previous chapters.

(2) Let $\tau \in (0,1)$ be the quantile level of interest for the estimated ZI GLMM based on the maximum likelihood estimate $\hat{\boldsymbol{\vartheta}}$. For each i, t, find $y_{it}^{\tau}|(\mathbf{x}_{it}, \mathbf{w}_{it}, \mathbf{z}_{it}) := q_{it}(\tau)$, which is the solution to $K_{it}(y) = 0$.

(3) Letting $\mathbf{y}_i^{\tau} = (y_{i1}^{\tau}, \ldots, y_{iT}^{\tau})^{\top}$, find the quantile-specific effects $\hat{\boldsymbol{\beta}}^{\tau}$ minimizing the risk function

$$\mathrm{E}\{L(\mathbf{y}_i^{\tau} - g(\hat{\boldsymbol{\eta}}_i))\}. \tag{3.9}$$

Note that $\hat{\boldsymbol{\eta}}_i$ is implicitly based on the maximum likelihood estimate $\hat{\boldsymbol{\vartheta}}$.

In order to obtain the estimated quantile effects $\hat{\boldsymbol{\beta}}^{\tau}$ in Step 3 above, we extend the NLS model in Chapter 2 to nonlinear mixed model (NLMM), accounting for the random intercepts from subject heterogeneity:

$$\begin{aligned} \mathbf{y}_i^{\tau} &= h(\boldsymbol{\eta}_i^{\tau}) + \boldsymbol{\epsilon}_i \\ \boldsymbol{\eta}_i^{\tau} &= \mathbf{X}_i\boldsymbol{\beta}^{\tau} + \mathbf{Z}_i\mathbf{u}_i^{\tau}, \end{aligned} \tag{3.10}$$

where $h(\cdot)$ is the same inverse log link function used for our ZI GLMMs and the $\boldsymbol{\epsilon}_i$ are $iid\ \mathcal{N}_T(\mathbf{0}, \sigma^2\mathbf{I}_T)$. Here, $\mathbf{I}_T$ is the $T \times T$ identity matrix. Notice that we find ourselves in the same situation as when performing MLE for the ZI GLMM. All of these quantities are explicitly written to show their dependency on the quantile τ since the NLMM being estimated has the conditional quantile $\mathbf{y}_i^{\tau}$ as the response.

Following the presentation in Chapter 7 of Pinheiro and Bates (2000), we first note that the variance-covariance matrix $\mathbf{G}^{\tau}$ of the random effects $\mathbf{u}^{\tau}$ can be rewritten in terms of the precision factor $\boldsymbol{\Delta}^{\tau}$, so that $(\mathbf{G}^{\tau})^{-1} = \sigma^{-2\tau}\boldsymbol{\Delta}^{\tau\top}\boldsymbol{\Delta}^{\tau}$. We further note that if $\mathbf{G}^{\tau} > 0$, as is assumed for our setting, then such a $\boldsymbol{\Delta}^{\tau}$ exists, but need not be unique. Estimation can be accomplished using penalized iteratively reweighted least squares using the two steps outlined in Lindstrom and Bates (1990): a penalized nonlinear least squares (PNLS) step and a linear mixed model (LMM) estimation step.

As in Chapter 2, we turn to the bootstrap for statistical inference for the three-step estimator and to provide confidence intervals in the empirical Section 3.7. The estimation is carried out by employing the multiplier bootstrap. Results presented in Section 3.4 found that the multiplier bootstrap confidence intervals' coverage probabilities tend to give results close to the nominal probabilities under the ZIP and the ZINB distributions. While this chapter focuses on identifying and estimating quantile functions for ZIP and ZINB distributions, we continue to investigate improvements in terms of statistical inference as, for example, exploring the use of bootstrap calibration to improve coverages of bootstrap-based confidence intervals (Loh, 1991).

3.4 Numerical Study

We next conduct a simulation study designed to evaluate our method's finite-sample performance proposed in Section 3.3. We first present results for models with a fixed proportion of zero inflation. Then we include simulations for the case where the model generates a proportion of zeros that varies by subjects and time.

We follow data generating processes similar to those considered in Machado and Santos Silva (2005), and extend them to the panel data setting. We consider that the response vector $\mathbf{y}_i$ for the i^{th} subject was generated from a count distribution subject to zero inflation. That is, y_{it} is generated from a degenerate distribution at zero with probability π_{it} and from a count distribution $p_{y_{it}|\mathbf{u}_i}$ with probability $1 - \pi_{it}$. In our numerical work, the Poisson and negative binomial portions of the ZI models each have the following conditional mean:

$$\mu_{it} = \exp\{\beta_0 + \beta_1 x_{it} + \beta_2 x_i + u_i\}$$

where $x_{it} = r_0 + r_1\epsilon_i + r_2\epsilon_{it}$, and the variables ϵ_i and ϵ_{it} are iid Gaussian random variables. The variable u_i is $iid\, \mathcal{N}(0, \sigma_u^2)$, where $\sigma_u^2 = 0.2$. The values for the time-invariant regressor x_i are chosen as equally-spaced design points over the interval $[0, 10]$. In all the simulation settings, $\beta_0 = 0.75$, $\beta_1 = r_1 = 0.25$, $\beta_2 = r_0 = 0$, and $r_2 = 1$.

We consider $N \in \{150, 250\}$ and $T \in \{5, 10\}$. The aforementioned simulation settings allow us to identify and estimate mean effects and quantile effects at $\tau \in \{0.50, 0.75, 0.90\}$. To evaluate the small sample performance of our approach, we report the bias and root mean square error (RMSE) of the first-step and third-step estimators. That is, the results for the parameters β_1 and β_2 of the conditional mean model estimated in the first step, and the results for coefficients β_1^τ and β_2^τ estimated using the quantile response variable, as a solution of the nonlinear equation specified in the third step. The bias and RMSE of the mean effect are calculated with respect to the parameter $\beta_1 = 0.25$ and $\beta_2 = 0$. However, we do not have a true parameter value of $\hat{\beta}(\tau)$ to which we can reference. The strategy we employ is to determine *pseudo-true parameter values* via simulation using very large N. Different n were also explored for the same data generating process. As can be seen from Table 3.1, the values have already stabilized at the given sample size; hence we will use the values at the sample size $N = 20,000$. Albeit our strategy to calculate the pseudo-true parameter values of the $\hat{\beta}(\tau)$'s is simulation-based, it is done so in a spirit similar to the notion of pseudo-true parameter values as defined in the context of model selection; see Sawa (1978) and Vuong (1989).

Table 3.1: Average slope values at each quantile level for the Poisson GLMM (top-half of table) and negative binomial GLMM (bottom-half of table). Results are based on $M = 5000$ simulations.

N	n	$\beta_1(0.50)$	$\beta_1(0.75)$	$\beta_1(0.90)$
Poisson				
10000	5	0.265	0.223	0.197
	10	0.265	0.223	0.197
	25	0.265	0.223	0.197
20000	5	0.265	0.223	0.197
	10	0.265	0.223	0.197
	25	0.265	0.223	0.197
Negative Binomial				
10000	5	0.268	0.238	0.225
	10	0.268	0.238	0.225
	25	0.268	0.238	0.225
20000	5	0.268	0.238	0.225
	10	0.268	0.238	0.225
	25	0.268	0.238	0.225

Table 3.2 and Table 3.3 show the small sample performance of the estimators for β_1 and β_2 when y_{it} is distributed as ZIP with a constant proportion of zero inflation. We consider $\pi_{it} \in \{0, 0.15, 0.30\}$, corresponding to no zero inflation, moderate zero inflation, and high zero inflation, respectively. The table shows that the methods perform quite well, yielding negligible biases for the mean effect and quantile effects. For each combination of subjects and time, the table shows small biases and excellent RMSE performance. Moreover, the table highlights that the proposed approach is robust to zero inflation. The performance of the proposed approach is stable as the proportion of zero inflation increases. The biases are almost unchanged, while the RMSEs only slightly increase for the respective cases. When we turn our attention to the case of ZINB shown in Table 3.4 and Table 3.5, we find similar results. The methods continue to perform well, and the results do not seem to vary across the different proportions of zero inflation.

Table 3.2: Bias and RMSE of β_1 estimators when the response is distributed as ZIP.

N	T	Method	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
$\pi_{it} = 0$								
150	5	TS	0.002	0.028	-0.001	0.024	-0.002	0.021
		JIT	0.003	0.039	0.019	0.040	0.025	0.044
150	10	TS	0.000	0.019	-0.001	0.016	-0.001	0.014
		JIT	0.003	0.027	0.018	0.031	0.025	0.037
250	5	TS	0.002	0.021	-0.001	0.018	-0.002	0.016
		JIT	0.004	0.029	0.019	0.032	0.025	0.038
250	10	TS	0.000	0.014	-0.001	0.012	-0.001	0.011
		JIT	0.004	0.022	0.019	0.027	0.025	0.033
$\pi_{it} = 0.15$								
150	5	TS	0.003	0.032	-0.001	0.027	-0.002	0.024
		JIT	0.028	0.060	0.027	0.048	0.029	0.049
150	10	TS	-0.001	0.022	-0.002	0.019	-0.003	0.017
		JIT	0.026	0.045	0.026	0.039	0.028	0.040
250	5	TS	0.003	0.025	-0.001	0.021	-0.002	0.018
		JIT	0.028	0.049	0.028	0.041	0.029	0.042
250	10	TS	0.001	0.017	-0.001	0.014	-0.001	0.013
		JIT	0.028	0.040	0.027	0.035	0.028	0.036
$\pi_{it} = 0.30$								
150	5	TS	0.003	0.036	-0.002	0.031	-0.003	0.027
		JIT	0.078	0.116	0.041	0.063	0.032	0.053
150	10	TS	0.001	0.024	-0.001	0.021	-0.002	0.019
		JIT	0.083	0.103	0.040	0.054	0.033	0.046
250	5	TS	0.004	0.028	-0.001	0.024	-0.003	0.021
		JIT	0.084	0.107	0.042	0.056	0.034	0.048
250	10	TS	0.002	0.019	-0.001	0.016	-0.001	0.014
		JIT	0.086	0.098	0.042	0.049	0.034	0.041

Table 3.3: Bias and RMSE of β_2 estimators when the response is distributed as ZIP.

N	T	Method	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
$\pi_{it} = 0$								
150	5	TS	0.000	0.016	0.000	0.013	0.000	0.012
		JIT	0.000	0.018	0.000	0.016	0.000	0.017
150	10	TS	0.001	0.015	0.001	0.012	0.001	0.011
		JIT	0.000	0.016	0.001	0.015	0.001	0.015
250	5	TS	0.000	0.013	0.000	0.011	0.000	0.009
		JIT	0.000	0.014	0.000	0.013	0.000	0.014
250	10	TS	0.000	0.011	0.000	0.009	0.000	0.008
		JIT	0.000	0.012	0.000	0.011	0.000	0.012
$\pi_{it} = 0.15$								
150	5	TS	-0.001	0.017	-0.001	0.014	-0.001	0.012
		JIT	-0.001	0.022	-0.001	0.017	-0.001	0.018
150	10	TS	0.000	0.016	0.000	0.013	0.000	0.011
		JIT	-0.001	0.019	-0.001	0.016	0.000	0.016
250	5	TS	0.000	0.014	0.000	0.011	0.000	0.010
		JIT	0.000	0.018	0.000	0.014	0.000	0.014
250	10	TS	0.000	0.012	0.000	0.010	0.000	0.009
		JIT	0.000	0.015	0.000	0.012	0.000	0.012
$\pi_{it} = 0.30$								
150	5	TS	0.000	0.019	0.000	0.016	0.000	0.014
		JIT	0.000	0.036	0.000	0.021	0.000	0.019
150	10	TS	0.000	0.017	0.000	0.014	0.000	0.012
		JIT	-0.001	0.028	0.000	0.017	0.000	0.017
250	5	TS	0.000	0.015	0.000	0.012	0.000	0.011
		JIT	0.000	0.028	0.000	0.016	0.000	0.015
250	10	TS	0.001	0.012	0.000	0.010	0.000	0.009
		JIT	0.001	0.021	0.001	0.013	0.000	0.012

Table 3.4: Bias and RMSE of β_1 estimators when the response is distributed as ZINB.

N	T	Method	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
$\pi_{it} = 0$								
150	5	TS	0.004	0.039	-0.001	0.036	-0.003	0.034
		JIT	0.002	0.052	0.012	0.048	0.012	0.052
150	10	TS	-0.001	0.028	-0.002	0.025	-0.003	0.024
		JIT	0.000	0.038	0.011	0.036	0.012	0.038
250	5	TS	0.004	0.031	-0.001	0.028	-0.003	0.027
		JIT	0.000	0.040	0.012	0.039	0.014	0.041
250	10	TS	-0.001	0.022	-0.001	0.020	-0.003	0.019
		JIT	0.000	0.029	0.012	0.029	0.012	0.031
$\pi_{it} = 0.15$								
150	5	TS	0.002	0.049	-0.004	0.045	-0.006	0.042
		JIT	0.024	0.085	0.017	0.060	0.013	0.058
150	10	TS	0.000	0.032	-0.002	0.030	-0.003	0.028
		JIT	0.026	0.062	0.019	0.044	0.014	0.042
250	5	TS	0.006	0.036	-0.001	0.033	-0.003	0.031
		JIT	0.027	0.067	0.020	0.047	0.016	0.045
250	10	TS	0.001	0.026	-0.001	0.024	-0.003	0.022
		JIT	0.026	0.051	0.020	0.037	0.015	0.034
$\pi_{it} = 0.30$								
150	5	TS	0.007	0.053	-0.001	0.049	-0.004	0.046
		JIT	0.044	0.117	0.034	0.077	0.020	0.064
150	10	TS	0.003	0.036	0.000	0.033	-0.002	0.031
		JIT	0.049	0.088	0.036	0.060	0.020	0.047
250	5	TS	0.007	0.041	-0.002	0.037	-0.005	0.036
		JIT	0.047	0.094	0.033	0.061	0.018	0.050
250	10	TS	0.003	0.029	-0.001	0.027	-0.003	0.026
		JIT	0.053	0.077	0.034	0.053	0.019	0.039

Table 3.5: Bias and RMSE of β_2 estimators when the response is distributed as ZINB.

N	T	Method	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
$\pi_{it} = 0$								
150	5	TS	-0.001	0.019	-0.001	0.017	-0.001	0.015
		JIT	-0.001	0.022	-0.001	0.020	-0.001	0.021
150	10	TS	0.000	0.016	0.000	0.014	0.000	0.013
		JIT	0.000	0.018	0.000	0.017	0.000	0.018
250	5	TS	0.000	0.014	0.000	0.013	0.000	0.012
		JIT	0.000	0.017	0.000	0.015	0.000	0.016
250	10	TS	0.001	0.012	0.001	0.011	0.001	0.010
		JIT	0.001	0.014	0.000	0.013	0.001	0.013
$\pi_{it} = 0.15$								
150	5	TS	0.000	0.021	0.000	0.018	0.000	0.017
		JIT	0.000	0.029	0.000	0.023	0.000	0.022
150	10	TS	-0.001	0.017	0.000	0.015	0.000	0.014
		JIT	-0.001	0.024	-0.001	0.019	-0.001	0.018
250	5	TS	-0.001	0.016	-0.001	0.014	0.000	0.013
		JIT	-0.001	0.022	-0.001	0.017	0.000	0.017
250	10	TS	0.000	0.013	0.000	0.012	0.000	0.011
		JIT	0.000	0.018	0.000	0.014	0.000	0.014
$\pi_{it} = 0.30$								
150	5	TS	0.001	0.023	0.001	0.020	0.001	0.018
		JIT	0.002	0.045	0.002	0.026	0.001	0.024
150	10	TS	0.000	0.019	0.000	0.016	0.000	0.015
		JIT	0.000	0.034	0.000	0.021	0.000	0.019
250	5	TS	0.000	0.018	0.000	0.016	0.000	0.014
		JIT	0.000	0.035	0.000	0.021	0.000	0.019
250	10	TS	0.000	0.014	0.000	0.012	0.000	0.012
		JIT	-0.001	0.026	0.000	0.016	0.000	0.015

Table 3.6: Bias and RMSE when the response is distributed as ZIP under a varying proportion of zeros.

N	T	Method	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
150	5	TS	0.002	0.032	-0.002	0.027	-0.003	0.024
		JIT	0.026	0.056	0.027	0.047	0.028	0.048
150	10	TS	0.001	0.021	-0.001	0.018	-0.001	0.016
		JIT	0.028	0.046	0.027	0.039	0.028	0.039
250	5	TS	0.004	0.025	-0.001	0.021	-0.002	0.018
		JIT	0.028	0.048	0.028	0.041	0.028	0.041
250	10	TS	0.001	0.016	-0.001	0.014	-0.001	0.013
		JIT	0.027	0.039	0.027	0.034	0.029	0.037

Next, we consider a scenario with varying proportions of zero inflation. The response variable y_{it} is now generated from the degenerate distribution at zero with probability π_{it} specified via the following logistic regression model:

$$\text{logit}(\pi_{it}) = \gamma_0 + \gamma_1 w_{it} \tag{3.11}$$

where w_{it} are $iid\ \mathcal{U}(0, 1)$ random variables and $\gamma_0 = -2$ and $\gamma_1 = 0.45$. This model specification generates a sequence of π_{it} that ranges over the interval $(0.12, 0.17)$ with an average of 0.145. In a first step not reported to save space, these probabilities are estimated using MLE.

Tables 3.6 and 3.8 show the small sample performance of the estimators for the slope parameters when y_{it} is distributed as ZIP and ZINB and π_{it} is generated as in equation (3.11). Once again, the tables show good performance of the estimator in each step. It also shows that the proposed approach is robust to different models of zero inflation, as indicated by the similar performance of the first-step and third-step estimators compared to their performance under constant zero inflation. The bias of the estimator is small, and the RMSE tends to decrease as the sample size increases, as expected.

These tables indicate that the model-aware approach yields negligible biases for the mean effect and the three quantile effects. For each combination of N (number of subjects) and n (number of replicates), the values of biases are on the same scale for each of the four estimators. A similar pattern is also observed for the corresponding RMSE values. This highlights our approach's empirical performance for both the mean structure and the quantile levels is practically equivalent.

Table 3.7: Bias and RMSE when the response is distributed as ZIP under a varying proportion of zeros.

N	T	Method	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
150	5	TS	0.000	0.017	0.000	0.014	0.000	0.012
		JIT	0.000	0.022	0.000	0.018	0.000	0.018
150	10	TS	-0.001	0.015	-0.001	0.013	-0.001	0.011
		JIT	-0.001	0.019	-0.001	0.015	-0.001	0.016
250	5	TS	0.000	0.014	0.000	0.011	0.000	0.010
		JIT	0.000	0.017	0.000	0.014	0.000	0.014
250	10	TS	0.000	0.012	0.000	0.010	0.000	0.009
		JIT	0.000	0.015	0.000	0.012	0.000	0.012

Table 3.8: Bias and RMSE for β_1 when the response is distributed as ZINB under a varying proportion of zeros.

N	T	Method	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
150	5	TS	0.005	0.046	-0.001	0.042	-0.003	0.040
		JIT	0.027	0.081	0.019	0.058	0.015	0.056
150	10	TS	0.000	0.031	-0.001	0.028	-0.003	0.027
		JIT	0.024	0.060	0.019	0.043	0.015	0.041
250	5	TS	0.005	0.036	-0.003	0.033	-0.005	0.032
		JIT	0.024	0.065	0.019	0.047	0.014	0.045
250	10	TS	0.001	0.025	-0.001	0.023	-0.002	0.021
		JIT	0.026	0.050	0.020	0.037	0.015	0.033

Table 3.9: Bias and RMSE for β_2 when the response is distributed as ZINB under a varying proportion of zeros.

N	T	Method	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
150	5	TS	0.000	0.021	0.000	0.018	0.000	0.017
		JIT	-0.001	0.030	0.000	0.023	0.000	0.022
150	10	TS	0.001	0.017	0.001	0.015	0.001	0.014
		JIT	0.001	0.023	0.001	0.019	0.001	0.019
250	5	TS	0.000	0.016	0.000	0.014	0.000	0.013
		JIT	0.001	0.023	0.000	0.017	0.000	0.017
250	10	TS	0.000	0.013	0.000	0.012	0.000	0.011
		JIT	0.000	0.018	0.000	0.014	0.000	0.014

Table 3.10: Bias and RMSE of two bootstrap routines when the response is distributed as ZIP.

N	T	Pair-Block		Multiplier	
		Bias	RMSE	Bias	RMSE
$\pi_{it} = 0$					
150	5	0.001	0.036	0.000	0.033
150	10	0.001	0.023	-0.001	0.022
250	5	0.002	0.027	-0.002	0.024
250	10	-0.001	0.018	-0.002	0.018
$\pi_{it} = 0.15$					
150	5	0.005	0.044	0.005	0.040
150	10	0.001	0.028	0.001	0.026
250	5	0.003	0.033	0.003	0.031
250	10	0.000	0.022	0.000	0.020
$\pi_{it} = 0.30$					
150	5	0.002	0.051	0.010	0.046
150	10	0.002	0.032	0.004	0.029
250	5	0.004	0.037	0.007	0.035
250	10	0.000	0.025	0.003	0.023

Moreover, these tables highlight that the model-aware approach is robust to zero inflation. The performance is stable as the proportion of zero inflation increases or changes. The biases are almost unchanged, while the RMSEs only slightly increase for the respective cases. This shows that one can reasonably expect to get meaningful estimates of $\beta_1(\tau)$ in the presence of zero inflation, thus allowing one to convey practical interpretations about the behavior of the sampled population at the τ^{th} quantile in the presence of zero inflation.

We finally turn to confidence interval estimation by employing the bootstrap. A comparison of two bootstrap implementations, pair-block bootstrap and multiplier bootstrap, reveals similar performance in Table 3.10 and Table 3.11. Hence, we focus on the multiplier bootstrap for confidence interval estimation.

Table 3.12 and 3.13 give the coverage probabilities for the 95% multiplier bootstrap confidence intervals under the ZIP GLMM and the ZINB GLMM, respectively. Confidence intervals are constructed by reporting the middle 95% of the multiplier bootstrap samples. However, the empirical coverage probabilities in these tables are liberal relative to the nominal level. This is noticeable as the amount of zero inflation increases, suggesting that improvements could be sought through, for example, a bootstrap calibration (Loh, 1991).

Table 3.11: Bias and RMSE of two bootstrap routines when the response is distributed as ZINB.

N	T	Pair-Block		Multiplier	
		Bias	RMSE	Bias	RMSE
		$\pi_{it} = 0$			
150	5	0.001	0.054	0.001	0.051
150	10	0.001	0.036	0.000	0.034
250	5	0.002	0.040	0.000	0.036
250	10	-0.001	0.028	-0.001	0.026
		$\pi_{it} = 0.15$			
150	5	0.001	0.064	0.004	0.058
150	10	-0.001	0.042	0.001	0.038
250	5	0.001	0.048	0.000	0.044
250	10	-0.001	0.033	0.000	0.029
		$\pi_{it} = 0.30$			
150	5	-0.001	0.074	0.005	0.067
150	10	-0.002	0.048	0.000	0.043
250	5	-0.002	0.054	0.000	0.049
250	10	-0.002	0.038	0.001	0.033

Table 3.12: Coverage probabilities for multiplier bootstrap confidence intervals for the mean effect and quantile effects. The data were generated from a ZIP GLMM with results based on $M = 400$ simulations with $B = 200$ bootstrap samples in each simulation.

n	m	$\hat{\beta}_1$(Mean)	$\hat{\beta}_1$(0.5)	$\hat{\beta}_1$(0.75)	$\hat{\beta}_1$(0.9)
		$p_{0,it} = 0$			
150	5	0.895	0.898	0.900	0.900
150	10	0.918	0.918	0.905	0.915
250	5	0.922	0.925	0.925	0.925
250	10	0.938	0.930	0.940	0.930
		$p_{0,it} = 0.15$			
150	5	0.862	0.872	0.868	0.872
150	10	0.908	0.918	0.908	0.912
250	5	0.900	0.902	0.895	0.895
250	10	0.918	0.922	0.918	0.915
		$p_{0,it} = 0.30$			
150	5	0.885	0.895	0.890	0.888
150	10	0.922	0.930	0.928	0.918
250	5	0.840	0.848	0.840	0.858
250	10	0.885	0.902	0.888	0.895

Table 3.13: Coverage probabilities for multiplier bootstrap confidence intervals for the mean effect and quantile effects. The data were generated from a ZINB GLMM with results based on $M = 400$ simulations with $B = 200$ bootstrap samples in each simulation.

n	m	$\hat{\beta}_1$(Mean)	$\hat{\beta}_1$(0.5)	$\hat{\beta}_1$(0.75)	$\hat{\beta}_1$(0.9)
$\pi_{it} = 0$					
150	5	0.882	0.870	0.875	0.852
150	10	0.882	0.890	0.885	0.872
250	5	0.865	0.870	0.858	0.845
250	10	0.900	0.910	0.910	0.892
$\pi_{it} = 0.15$					
150	5	0.852	0.858	0.848	0.840
150	10	0.890	0.895	0.908	0.898
250	5	0.895	0.888	0.882	0.870
250	10	0.862	0.865	0.862	0.862
$\pi_{it} = 0.30$					
150	5	0.872	0.872	0.875	0.855
150	10	0.868	0.872	0.860	0.862
250	5	0.902	0.888	0.892	0.878
250	10	0.892	0.895	0.898	0.885

3.5 An Application using the RAND Health Insurance Experiment

Using data from Deb and Trivedi (2002), this section investigates how medical care utilization measured by the number of visits to a medical doctor (MD) is affected by health insurance plans, demographic characteristics, and health status of patients. Over 30% of the observations are zeros, motivating the use of the proposed approach. From a health policy viewpoint, it is important to understand how policies affect the participants who need health care. Hence, a distinction between non-users and users helps learn the effect of policy more precisely. Overall, the conditional quantile functions and effects reported in this section contribute to an informative discussion that goes beyond mean effects. We find that the effect of insurance variables and demographics vary across the conditional distribution of medical care utilization, while revealing interesting differences with respect to results obtained by existing methods that ignore subject heterogeneity and zero inflation.

Data

In the 1970s, the RAND Corporation initiated the 15-year, multimillion-dollar social experiment in health care research. This remains the largest and longest controlled experiment on health policy in U.S. history. The RAND Health Insurance Experiment (RHIE) was originally designed to study how multiple factors affected the usage of medical care and the corresponding participants' health consequences. During the study, data were collected from participants of 2823 families, where each family was enrolled in the insurance plans for 3 or 5 years.

In this paper, we analyzed one subset of data from the RHIE as in Deb and Trivedi (2002), where the participants were only enrolled in the fee-for-service plans. This particular dataset consists of 5908 participants with 20,186 observations in total: each participant has 3 or 5 observations; each observation corresponds to data collected for the participant in a given year. The response variable MDU is the yearly count of outpatient visits to physicians, which represents the health care utilization for the experimental subject for a specific year. The insurance variables were randomly assigned and include a coinsurance rate (LC), an indicator variable for plans with a deductible (IDP), a maximum dollar-expenditure function (FMDE), and a participation-incentive payment function (LPI). Other covariates include factors representing the participants' socioeconomic status, demographic information, and health status. For detailed variable definitions and summary statistics, see Table 3.17 in the Appendix.

The use of a ZI count model is also supported by some features of the RHIE. As mentioned in Section 3.1, the distribution of the response shows medium-to-high proportion of zero utilization. Moreover, while a number of people are healthy during the period and they have no need to visit hospitals at all, a number of patients are unhealthy and have the need to visit physicians. Depending on the severeness and the practical considerations (for example, the possible payment to the health care service), some patients might not go to the physicians while others have multiple visits. Under this circumstance, a random zero count could be observed but a positive integer-valued count is also possible.

Model specification

In the first step, we model the conditional mean considering four different specifications: Poisson model, negative binomial model, ZIP model, and ZINB model. Due to the existence of both zero inflation and a long tail to the right, the ZINB model provides the best fit. This is supported by the evidence from the information criterion and an assessment of the randomized quantile residuals (Dunn and Smyth, 1996).

We considered four count regression models: Poisson regression model, ZIP regression model, negative binomial regression model, and ZINB regression model. Model comparisons via the AIC values are consistent with that based on BIC values; results are provided in Table (3.14). Table (3.14) shows that the ZINB regression models provide a better fit.

Table 3.14: Model comparisons via AIC/BIC values with degree of freedoms (df) in the parentheses.

Model	AIC(df)	BIC(df)
Poisson Regression	85757.14(19)	85907.49(19)
ZIP Regression	84418.78(20)	84577.04(20)
NB Regression	80763.38(20)	80921.63(20)
ZINB Regression	80548.38(26)	80754.11(26)

The randomized quantile residuals under each model provide a visual illustration for GOF assessment. A comparison reveals huge differences among models. Clearly, the Poisson regression model shows the worst fit because it fails to capture the presence of zero-inflation and certain large values. ZIP regression shows some improvements over Poisson regression but still fails to tackle the overdispersion. On the other hand, both negative binomial regression and ZINB regression provide a satisfactory fit to the data, as indicated by the randomized quantile residual plot in Figure 3.2.

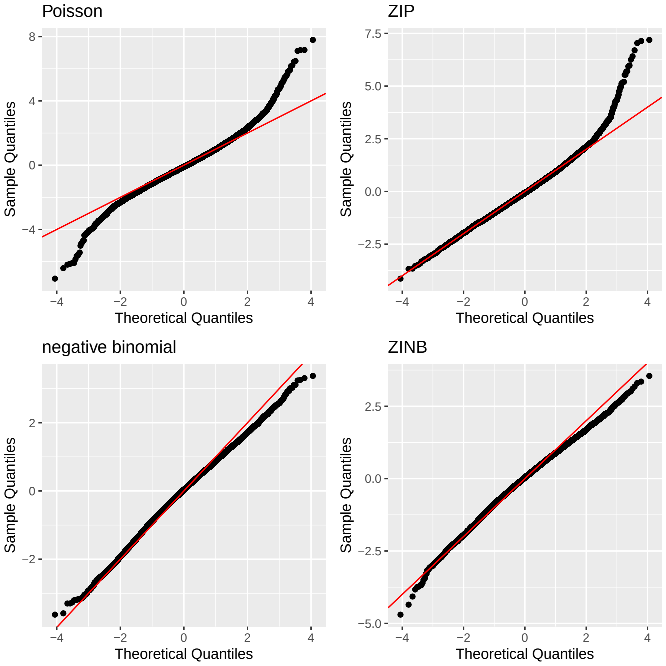


Figure 3.2: Randomized quantile residuals of fitted models. First row shows results for Poisson regression (left) and negative binomial regression (right); second row shows results for ZIP regression (left) and ZINB (right).

To test for the presence of zero-inflation, we conducted a boundary-corrected LR test. Results for three tests are reported in Table 3.15. All the tests show evidence for the presence of zero-inflation, and ZINB regression gives the best fit based on these results.

Since all the evidence supports the ZINB regression model, we will focus on the ZINB structure hereafter. We then estimate a ZINB count model for the number of visits to a medical doctor considering the vector of treatment variables and covariates used by Deb and Trivedi (2002). We estimate π_{it} as a function of a covariate vector $\boldsymbol{w}_{it}$ that includes

Table 3.15: Results of the boundary likelihood ratio test. All p-values are significant at 0.001

Models tested	Test statistic	Results
ZIP *versus* Poisson	1340.4	ZIP
ZINB *versus* negative binomial	227.0	ZINB
ZINB *versus* ZIP	3882.4	ZINB

Table 3.16: Estimated regression coefficients for the RAND Health Insurance Experiment dataset.

Variables	Mean	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$
LC	-0.059 (0.020)	-0.109 (0.006)	-0.051 (0.006)	-0.046 (0.007)
IDP	-0.165 (0.038)	-0.281 (0.013)	-0.151 (0.013)	-0.142 (0.017)
LPI	0.013 (0.006)	0.024 (0.002)	0.012 (0.002)	0.011 (0.002)
FMDE	-0.020 (0.012)	-0.026 (0.004)	-0.019 (0.003)	-0.018 (0.003)
LINC	0.079 (0.014)	0.140 (0.005)	0.085 (0.003)	0.063 (0.004)
LFAM	-0.125 (0.028)	-0.126 (0.019)	-0.119 (0.012)	-0.111 (0.011)
AGE	0.001 (0.001)	0.001 (0.001)	0.000 (0.001)	0.001 (0.001)
FEMALE	0.411 (0.036)	0.569 (0.017)	0.377 (0.020)	0.343 (0.020)
CHILD	0.359 (0.050)	0.390 (0.052)	0.340 (0.045)	0.316 (0.039)
FEMCHILD	-0.383 (0.053)	-0.395 (0.029)	-0.369 (0.029)	-0.339 (0.025)
BLACK	-0.538 (0.050)	-1.160 (0.018)	-0.490 (0.014)	-0.438 (0.022)
EDUCDEC	0.023 (0.006)	0.042 (0.002)	0.020 (0.002)	0.019 (0.002)
PHYSLIM	0.297 (0.046)	0.427 (0.023)	0.275 (0.017)	0.243 (0.018)
NDISEASE	0.028 (0.002)	0.044 (0.001)	0.025 (0.001)	0.023 (0.001)
HLTHG	0.019 (0.031)	0.013 (0.015)	0.020 (0.011)	0.014 (0.009)
HLTHF	0.208 (0.058)	0.249 (0.018)	0.194 (0.017)	0.177 (0.018)
HLTHP	0.537 (0.115)	0.802 (0.042)	0.514 (0.041)	0.448 (0.038)

LC, LPI, an indicator for children under the age of 18, an indicator for race, and the number of years of education of the head of the household. Moreover, the conditional mean and conditional quantile functions are augmented by individual-specific intercepts. Modeling individual-specific intercepts as random effects is consistent with the use of experimental data. A vital advantage of the RHIE data is that insurance plans were randomly assigned, and, consequently, the treatment variables are not correlated with individual-specific latent characteristics.

Empirical Results

Table 3.16 presents the conditional mean effects and quantile effects corresponding to insurance, demographics, and health status parameters. The first column presents point estimates for the mean parameters, and the last three columns show results for the quantile parameters estimated at the 0.5, 0.75, and 0.90 quantiles. The table also includes standard errors obtained by employing the bootstrap.

The point estimates corresponding to the first step shown in column 1 of Table 3.16 is similar to the estimates in Table 4 in Deb and Trivedi (2002). However, the estimates presented here are estimated more precisely. The sign of the coefficients is consistent with expectations and standard economic theory. For instance, the coefficient of LC can be interpreted as a price effect, and it is negative and significant at standard levels. We expect the effect of LC on the count variable to be negative because the patient's cost is higher as the rate of coinsurance increases. Also, as expected, the number of visits to an MD increases with the natural logarithm of income (LINC).

When we examine the effects across quantiles, we see some interesting differences in LC, LINC, and the indicator for race of the head of the household (BLACK). We find that the mean effect of these variables is quantitatively similar to the estimated effects at the 0.75 and 0.90 quantiles, revealing the importance of distributional effects and that the mean effect offers an incomplete description of the effect of some insurance, demographics, and socioeconomic variables. To examine this claim in more detail, we estimate the model as in Table 3.16 considering now 13 equally spaced quantiles τ in the interval $[0.3, 0.9]$. We then concentrate our attention on some of the variables considered in Table 3.16.

Figure 3.3, Figure 3.4 and Figure 3.5 illustrate the estimated mean effects (dashed lines) and quantile effects (continuous lines) obtained from our proposed method, across all selected quantiles. In order to examine the importance of accommodating for the large number of zeros in the RHIE, we also report estimates obtained by the jittering approach of Machado and Santos Silva (2005) and Harding and Lamarche (2019a). The figure shows some interesting new findings. First, we find that the health insurance option associated with coinsurance (LC) significantly reduce medical care utilization, particularly among

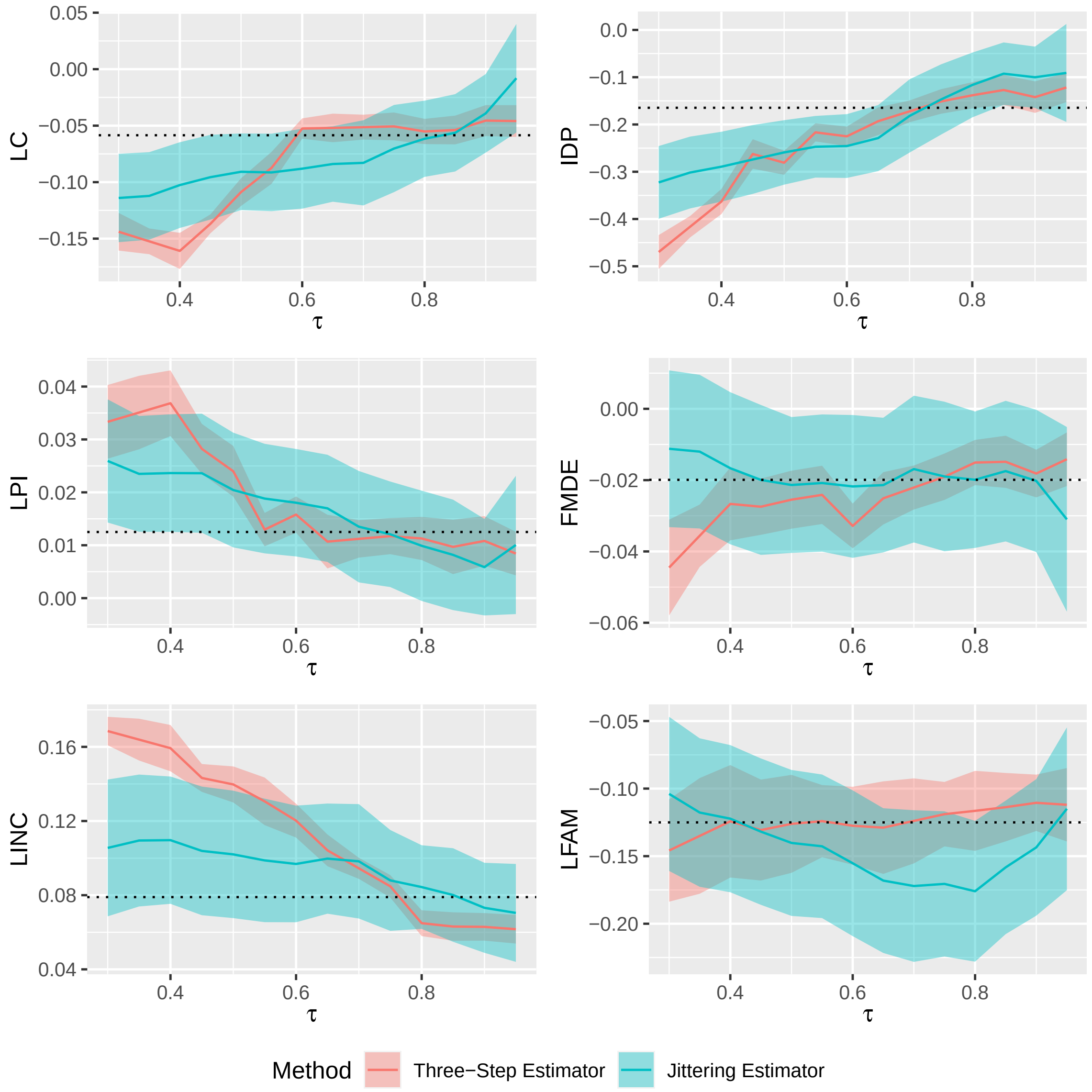


Figure 3.3: Estimated regression coefficients, $\hat{\boldsymbol{\beta}}^{\tau}$, for policy variables, socio-economic variables and demographic variables. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

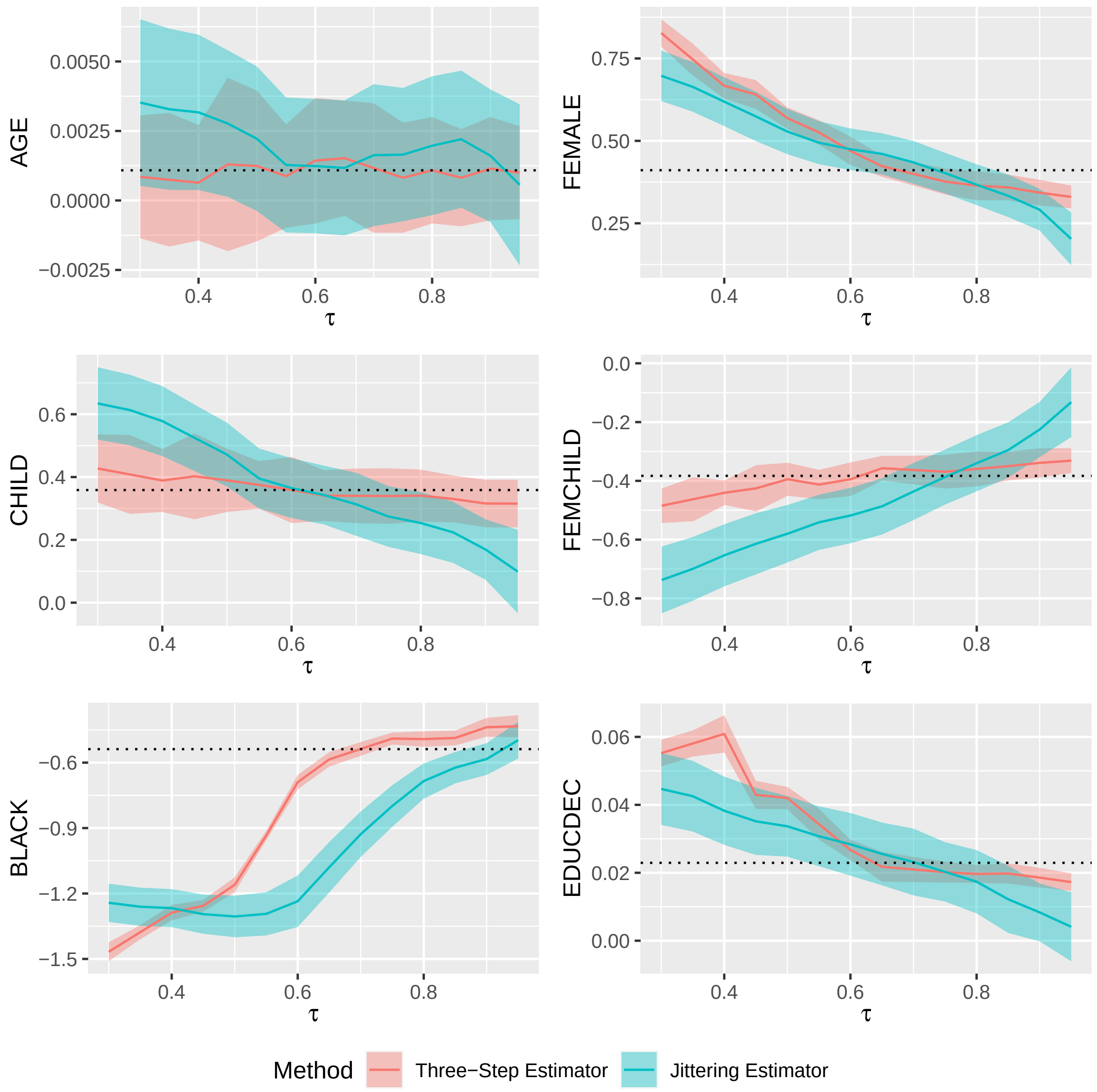


Figure 3.4: Estimated regression coefficients, $\hat{\boldsymbol{\beta}}^{\tau}$, for policy variables, socio-economic variables and demographic variables. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

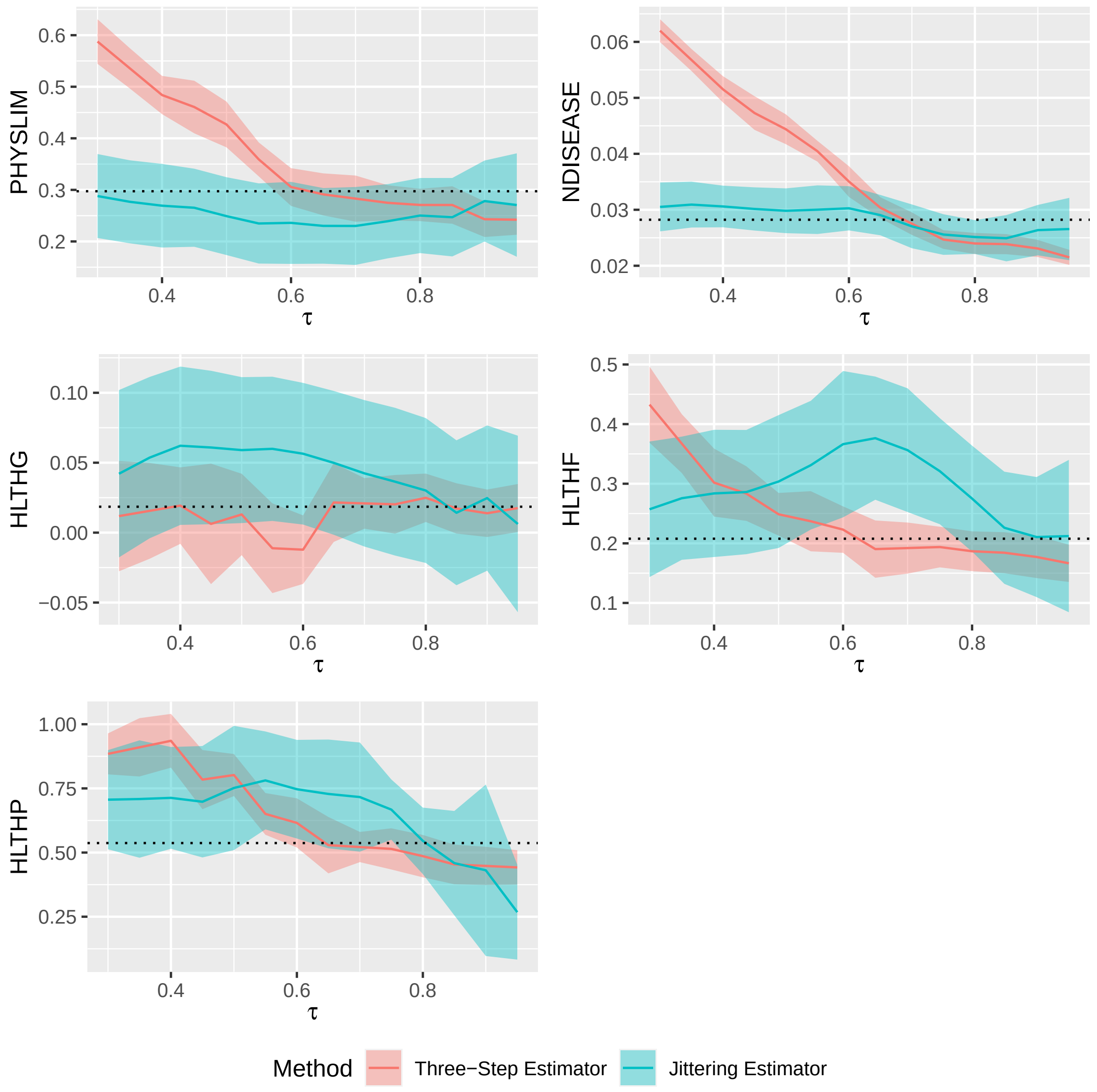


Figure 3.5: Estimated regression coefficients, $\hat{\boldsymbol{\beta}}^{\tau}$, for policy variables, socio-economic variables and demographic variables. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

those with conditionally low number of visits to a medical doctor. While the effect at the mean is -0.05, the effect at the 0.3 quantile is about three times smaller, revealing increasing price sensitivity at the lower tail. Interestingly, we also find a significant black-white gap in terms of utilization, and the estimates reveal that the gap widens as we move from the center of the conditional distribution to the lowest quantiles. Lastly, the comparison of the quantile effects for LINC and BLACK obtained by different methods reveal non-negligible differences arising from simultaneously addressing subject heterogeneity and zero inflation.

3.6 Discussion

The primary aim of this work is to study the identification and estimation of conditional quantile functions for discrete responses with zero inflation in the longitudinal data setting. Our approach uses a continuous approximation to the discrete distribution for the count model under consideration. This approach has been developed in Ilienko (2013b) and Padellini and Rue (2019a), with the latter leveraging this approximation to perform quantile regression for discrete data. We extended to the longitudinal setting where the count responses are also subject to zero inflation. Another important distinction from Padellini and Rue (2019a) is that we first consider estimation of the conditional mean rather than considering a quantile regression model. This critical innovation allows consistent estimation of a class of models with subject heterogeneity, without restrictions on the minimum number of repeated observations per subject.

The class of models used in our first step is ZI GLMMs, which affords the practitioner considerable flexibility regarding the structural form of the model for their application. The class of ZI GLMMs is, of course, predicated on classic GLMs, which formally require the dependent variable to be from a distribution in the exponential family. However, the ZI GLMMs are much broader in that one can model the parameters in a ZI model (including the mixing proportion) as a function of covariates. This includes distributions that are not part of the exponential family, like the negative binomial with an unknown dispersion parameter. Having this broad class of distributions at our disposal allows for practical exploration of reasonable and meaningful structures to consider for the conditional mean structure for the application at hand. MLE is accomplished using the Laplace approximation to calculate the marginal likelihood, which can be performed using the `R` package `glmmTMB` (Brooks et al., 2017).

The BLUEs and BLUPs from our estimated ZI GLMM are used in our second step to obtain a conditional quantile variate as a solution of a nonlinear moment condition defined for the conditional mean. The material presented in Section 3.2 shows that the solution exists and is unique. Then, a flexible NLMM is employed for a model of conditional quantile responses. We demonstrated through extensive simulation work in Section 3.4 that the proposed estimator has satisfactory performance to estimate quantile effects under different degrees of zero inflation.

The efficacy of our procedure is highlighted by analyzing data from the RAND Health Insurance Experiment. While these data have been analyzed in the literature using count regression models, we have provided a thorough examination of quantile effects while capturing subject heterogeneity and the fact that the data are longitudinal and subject to zero inflation. Our analysis provides a more nuanced view that can inform health policy experts to understand how specific policies affect the participants who need health care. Overall, the empirical results obtained for this data analysis, combined with the extensive simulation results, suggest the benefit of our more sophisticated techniques to understand quantile effects when modeling ZI longitudinal count responses.

3.7 Appendix

Table 3.17: Variable definitions and summary statistics for the RAND Health Insurance Experiment dataset.

Variables	Definition	Mean	Min	25% Quantile	Median	75% Quantile	Max
MDU	Yearly number of outpatient visits to physicians	2.861	0	0	1	4	77
LC	ln(coinsurance+1), $0 \leq$ coinsurance rate ≤ 100	2.384	0	0.000	3.258	4.564	4.564
IDP	Indicator for individual deductible plan	0.260	0	0	0	1	1
LPI	ln(max(1,annual participation incentive payment))	4.709	0	4.064	6.109	6.620	7.164
FMDE	log(max(medical deductible expenditure))	4.030	0	0.000	6.095	6.959	8.294
PHYSLIM	Indicator for physical limitations	0.124	0	0	0	0	1
NDISEASE	Index of chronic diseases	11.244	0	6.900	10.576	13.732	58.600
LINC	ln(annual family income) in US dollars	8.708	0	8.582	8.984	9.257	10.283
LFAM	ln(family size)	1.248	0	1.099	1.386	1.609	2.639
EDUCDEC	Educations of household decision-makers in years	11.967	0	11	12	13	25
AGE	Age in years	25.718	0	11.462	24.195	37.402	64.275
FEMALE	Indicator for female	0.517	0	0	1	1	1
CHILD	Indicator for age less than 18	0.401	0	0	0	1	1
FEMCHILD	FEMALE*CHILD	0.194	0	0	0	0	1
BLACK	If race of household head is black: 1	0.182	0	0	0	0	1
HLTHG	If self-rated health is good: 1	0.362	0	0	0	1	1
HLTHF	If self-rated health is fair: 1	0.077	0	0	0	0	1
HLTHP	If self-rated health is poor: 1	0.015	0	0	0	0	1

Chapter 4 COMBINE Study

4.1 Introduction

As highlighted in the preceding chapters, we proposed a novel modeling strategy for characterizing the conditional quantiles of count data. Our three-step method is able to handle both zero-inflation and over-dispersion in the data; furthermore, the method is also applicable in longitudinal settings.

In this chapter, we summarize our method with an illustrative example for alcohol dependence treatment. This application uses data from the Combined Pharmacotherapies and Behavioral Interventions (COMBINE) Study. The main goal is to investigate how combinations of medical therapies and behavioral interventions affect alcohol dependency, measured by average number of daily drinks, while accounting for demographic and socioeconomic variables. In particular, it is important to understand how these treatments affect patients with different degrees of alcohol dependency, and how these effects vary over time. Hence, a quantile count regression model for longitudinal data can help characterize the treatment effects more precisely.

4.2 Data

From January 2001 to January 2004, the National Institute on Alcohol Abuse and Alcoholism sponsored a randomized controlled trial across 11 US academic sites. The main goal was to study whether the efficacy of certain medications for alcoholism can be improved with specialist intervention. Two medications, naltrexone and acamprosate, were used for medical therapy. Certified alcoholism treatment specialists were employed to provide specialist care and intervention.

A total of 1383 volunteers, all of whom were diagnosed with primary alcohol dependence, were admitted into this trial. They were randomly assigned to one of nine groups to receive one combination of treatments over 16 weeks. Eight of these nine groups received some medications, while the ninth group did not. The medication is one of placebo pills, naltrexone, acamprosate or both drug. One set of these four groups received additional specialist treatment known as combined behavioral intervention (CBI), while the other four groups only took the respective medications. The ninth group, which did not receive medication, only received CBI during the 16 weeks. Information were collected at the initial date before the treatment (week 0), during the treatment (week 8 and week 16), 6 months after the initial date (week 26), 12 months after the initial date (week 52) and 12 months after the conclusion of the treatment (week 68). For detailed description of the study design and assessment, see Anton et al. (2006).

In this chapter, we analyze a specific subset of data from the COMBINE study. This particular dataset consists of 1240 subjects with 5734 observations in total. Although the original data has 6 observations for each subject, some observations contain missing records or errors. Hence, the dataset we analyze contains less than 6 observations for some subjects. The response variable (MEAN) is the average number of daily drinks, which is treated as a count. One standard drink is defined to be 0.5 oz of absolute alcohol, or 10 oz of beer, or 4 oz of wine, or 1.0 oz of 100-proof liquor (Anton et al., 2006).

Originally, the main predictors of interest are the categorical variables for the nine treatments. However, evidence from existing literature (Anton et al., 2006; Greenfield et al., 2010; Zweben et al., 2008) found no difference across these nine treatment groups. Hence, to focus on the illustration of our application, we collapsed the 9 groups into 4 groups: the first group (Treatment 0) is the same as the original group 1, in which participants only received placebo pills without CBI; the next group (Treatment 1) includes the previous group 2-4, in which participants received some medications without CBI; the original group 5-8 became a new group (Treatment 2), in which participants received both medications and CBI; the last group (Treatment 3) is the same as the original group 9 (CBI-only). Furthermore, an indicator variable for post-treatment effect was included in the analysis. The “post-treatment” corresponds to measurements in week 26, 52 and 68, as opposed to measurements before and in week 16.

Other predictor variables include demographic information for gender(FEMALE), an indicator variable for marital status (Married) and Race/Ethnicity. Socioeconomic covariates include the years of education one received (EduYear) and an indicator variable for the

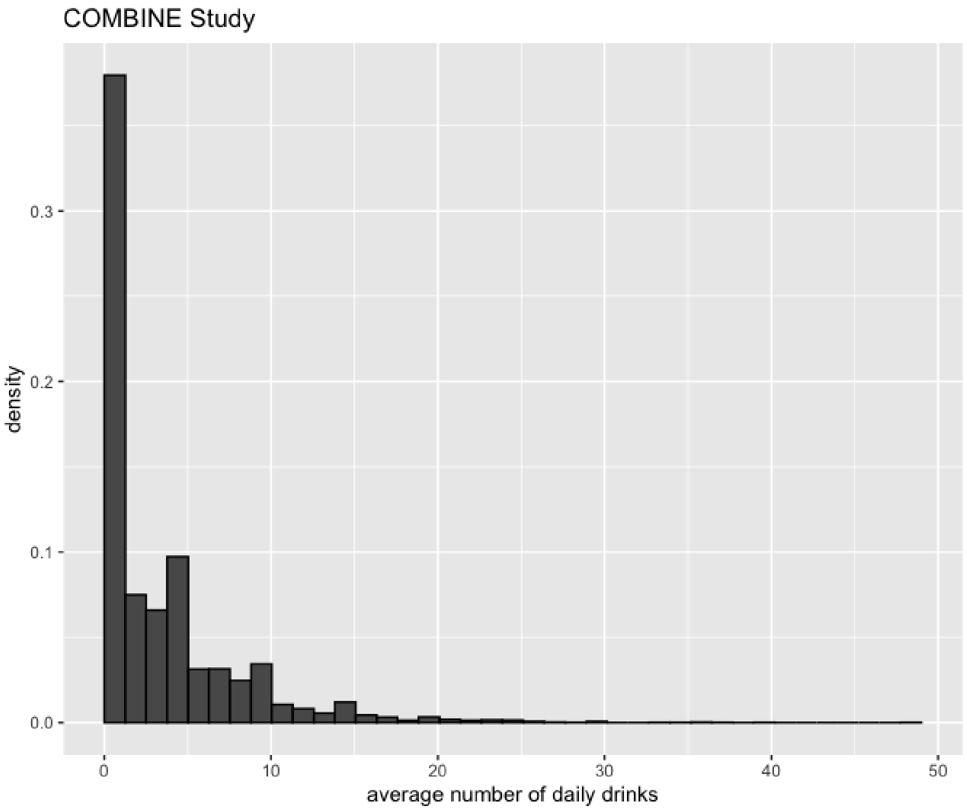


Figure 4.1: Average number of daily alcohol drinks in the COMBINE study data.

current working status (WORK). For detailed variable definitions and summary statistics, see Table 4.5 in the Appendix.

The proportion of zero drinks per day exceeds 35% for this particular dataset. Moreover, the count response distribution has a long tail extending to over 40 drinks per day. While the median is 2 drinks per day, the mean is 3.39. As in previous chapters, all these exploratory results show the need for a flexible approach that addresses zero-inflation, overdispersion and subject heterogeneity.

The use of a ZI count model is obvious from Figure 4.1 and exploratory analysis. As mentioned in Section 4.1, the distribution of the response shows medium-to-high proportion of zero drinks per day. Moreover, while some treatments might be highly efficient for some subjects during specific periods, such that these people are not drinking at all during the period, a number of patients are unhealthy and might relapse. Depending on the severeness of their addictions to alcohol and the received treatments, some patients might not want to drink while others have multiple drinks per day. Under this circumstance, a random zero count could be observed, but a non-zero count is also possible This is also supported by the fact that all of the COMBINE Study participants are alcoholics.

4.3 Model specification

In the first step, we model the conditional mean in a similar fashion. Four different specifications, Poisson model, negative binomial model, ZIP model, and ZINB model, are fitted to the data. Due to the existence of both zero inflation and a long tail to the right, the ZINB model provides the best fit. This is further supported by BIC and randomized quantile residuals (Dunn and Smyth, 1996).

We considered four count regression models: Poisson regression model, ZIP regression model, negative binomial regression model, and ZINB regression model. Model comparisons via the AIC values are consistent with those based on BIC values; results are provided in Table 4.1. Table 4.1 shows that the ZINB regression models provide a considerably better fit.

Table 4.1: Model comparisons via AIC/BIC values with degree of freedoms (df) in the parentheses.

Model	AIC(df)	BIC(df)
Poisson Regression	31544.41(15)	31644.22(15)
ZIP Regression	26930.53(18)	27050.31(18)
NB Regression	25942.72(16)	26049.19(16)
ZINB Regression	25654.54(19)	25780.97(19)

The randomized quantile residuals under each model provide a visual illustration for GOF assessment. A comparison reveals huge differences between models. Clearly, the Poisson regression model shows the worst fit because it fails to capture the presence of zero-inflation and certain large values. ZIP regression shows some improvements over Poisson regression but still fails to tackle the overdispersion. On the other hand, both negative binomial regression and ZINB regression provide a satisfactory fit to the data, as indicated by the randomized quantile residual plot in Figure 4.2.

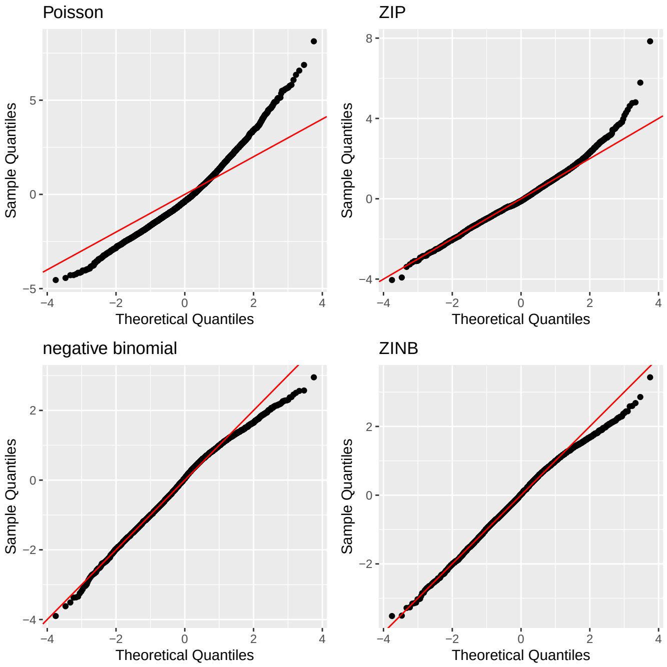


Figure 4.2: Randomized quantile residuals of fitted models. First row shows results for Poisson regression (left) and negative binomial regression (right); second row shows results for ZIP regression (left) and ZINB (right).

To test for the presence of zero-inflation, we conducted a boundary-corrected LR test. Results for three tests are reported in Table 4.2. All the tests show evidence for the presence of zero-inflation, and ZINB regression gives the best fit based on these results.

Table 4.2: Results of the boundary likelihood ratio test. All p-values are significant at 0.001

Models tested	Test statistic	Results
ZIP *versus* Poisson	4619.8	ZIP
ZINB *versus* negative binomial	294.2	ZINB
ZINB *versus* ZIP	1278	ZINB

Since all the evidence indicate that the ZINB regression model provides the best fit, we focus on the ZINB structure for the following steps. We modeled the average number of daily drinks with the vector of treatment variables plus some covariates used by Anton et al. (2006). We estimate π_{it} as a function of a covariate vector $\boldsymbol{w}_{it}$ that includes gender

Table 4.3: Estimated slopes and standard errors for the ZINB model (mean structure).

Variables	Estimates(SE)
Female	-0.469(0.050)
Married	-0.102(0.044)
Race(Black)	-0.169(0.088)
Race(Other)	0.175(0.108)
EduYear	-0.043(0.008)
Treatment1	-0.085(0.079)
Treatment2	-0.142(0.076)
Treatment3	-0.019(0.097)
Work	-0.373(0.050)
PostTreatment	-0.160(0.079)
Treatment1 × PostTreatment	-0.189(0.093)
Treatment2 × PostTreatment	-0.169(0.089)
Treatment3 × PostTreatment	-0.180(0.116)

(Female), current employment status (Work) and an indicator for measurements taken after the treatment period (Week 26, 52 and 68). Existing research (Edwards and Gross, 1976; Anton et al., 2006) on alcohol dependency has found evidence that females are less likely to become alcoholic, and that subjects with stable employment are less prone to alcohol dependency. Hence, it is reasonable to include the indicator variables for gender and current employment status for the logistic regression part. Also, patients show different drinking patterns after receiving these treatment combinations. As a result, the indicator variable for post-treatment is included for the model.

Table 4.4: Estimated regression coefficients for the COMBINE Study data.

Variables	Estimates(SE) at τ			
	0.30	0.50	0.70	0.90
Female	-0.585(0.043)	-0.496(0.045)	-0.451(0.044)	-0.413(0.068)
Married	-0.127(0.046)	-0.110(0.043)	-0.096(0.040)	-0.093(0.065)
Race(Black)	-0.235(0.103)	-0.179(0.093))	-0.160(0.110)	-0.157(0.134)
Race(Other)	0.213(0.163)	0.183(0.094)	0.180(0.121)	0.157(0.196)
EduYear	-0.053(0.008)	-0.046(0.007)	-0.041(0.008)	-0.038(0.013)
Treatment1	-0.108(0.094)	-0.102(0.094)	-0.089(0.075)	-0.073(0.158)
Treatment2	-0.166(0.081)	-0.162(0.082)	-0.143(0.068)	-0.122(0.149)
Treatment3	-0.032(0.102)	-0.018(0.094)	-0.020(0.086)	-0.009(0.211)
Work	-0.421(0.076)	-0.382(0.061)	-0.363(0.054)	-0.348(0.126)
PostTreatment	-0.178(0.099)	-0.180(0.064)	-0.160(0.064)	-0.147(0.214)
Treatment1 × PostTreatment	-0.217(0.120)	-0.179(0.085)	-0.176(0.076)	-0.177(0.233)
Treatment2 × PostTreatment	-0.211(0.112)	-0.160(0.073)	-0.160(0.068)	-0.156(0.221)
Treatment3 × PostTreatment	-0.244(0.129)	-0.184(0.118)	-0.183(0.093)	-0.169(0.316)

4.4 Empirical Results

Table 4.3 presents the conditional mean effects. At the mean level, females and married participants tend to drink less on a daily basis. The race and ethnicity show some difference across groups, however, these differences are not statistically significant. Education and current employment status are very significant predictors for the average number of drinks per day, which agrees with current research (Anton et al., 2006; Greenfield et al., 2010); more specifically, participants who have received longer years of education and who are currently working are drinking less after receiving the treatment. The treatments and their interactions with post-treatment indicator all have negative signs. This indicates that the therapies show some efficacy in reducing the alcohol dependency, although the efficacy is not strong. One particular treatment combination, Treatment 2 in our notation, shows the greatest efficacy in reducing daily drinks and is highly significant in our model. This treatment combination includes the three groups that received some medication therapy (naltrexone, acamprosate or both) and CBI. This is similar to the findings in Anton et al. (2006) and emphasizes the importance of combining medication therapy with behavioral intervention for the best results.

Table 4.4 presents the conditional quantile effects. The table presents point estimates for the selected predictors with SEs in the parentheses. SEs were obtained by the multiplier bootstrap with 200 bootstrap samples. Table 4.4 includes seven equally spaced quantiles τ in the interval $[0.3, 0.9]$

The point estimates corresponding to the third step show some changes compared to the estimates in Table 4.3. However, the estimates presented here are generally close to the estimates at the mean level. The signs of the coefficients are the same and the ranges of the values are consistent. This is uncommon from a QR viewpoint, where quantile estimates usually change as the quantile level τ changes. The exceptions are gender variable (Female), years of completed education (EduYear) and current employment status (Work), as we can see some stable changes across the quantiles. The estimated slopes for these three predictors are all negative, suggesting similar results as for the mean level. For example, females tend to drink less per day. Their effects are greater at the lower quantile than at the higher quantile; that is, as the quantile level τ increases, the estimated regression coefficients move towards zero. In other words, their effects diminish for alcoholics who drink considerably more. This calls for attention to patients who drink heavily on a daily basis, as heavy alcohol dependency is dangerous and difficult to treat.

It is also worth noting that these three variables are the most significant predictors for the mean model, where as other predictors show small-to-none significance. This finding is interesting as in the QR settings, variables might show different patterns compared to those in the mean regression setting. The fact that the mean effect of these variables is quantitatively similar to the estimated effects at the quantiles indicates the minimal importance of the variables. This holds not only for the average, but also for the complete distribution.

Figure 4.3 illustrate the estimated mean effects (dashed lines) and quantile effects (continuous lines) for the demographic variables. Figure 4.4 illustrate the estimated mean effects (dashed lines) and quantile effects (continuous lines) for the socio-economic variables. Figure 4.5 illustrate the estimated mean effects (dashed lines) and quantile effects (continuous lines) for the therapy variables and Figure 4.3 illustrate those for the interaction terms. The figures provide a visual illustration of Table 4.3 and Table 4.4. As we discussed, we find that the gender, years of completed education and current working status reduce daily number of drinks, particularly among those with slight alcohol consumption. This suggests the importance of prevention for heavy alcohol dependence, as policies tend to work the best when patients are not drinking too much. The same conclusion also holds for other variables, although the effects for the heavy drinkers seem to be similar to that of the rectified patients.

4.5 Discussion

In this chapter, we provided an empirical case study using data from the COMBINE Study. The purpose is to illustrate our method for count data with zero-inflation and subject heterogeneity. The analysis helps provide new insights into alcohol dependency and therapies.

Since the response variable is a count, the use of a count regression model is suitable. Furthermore, we illustrate our modeling strategy that includes variable selection, model comparison and diagnostics. These provide valuable assessments for the validity of the

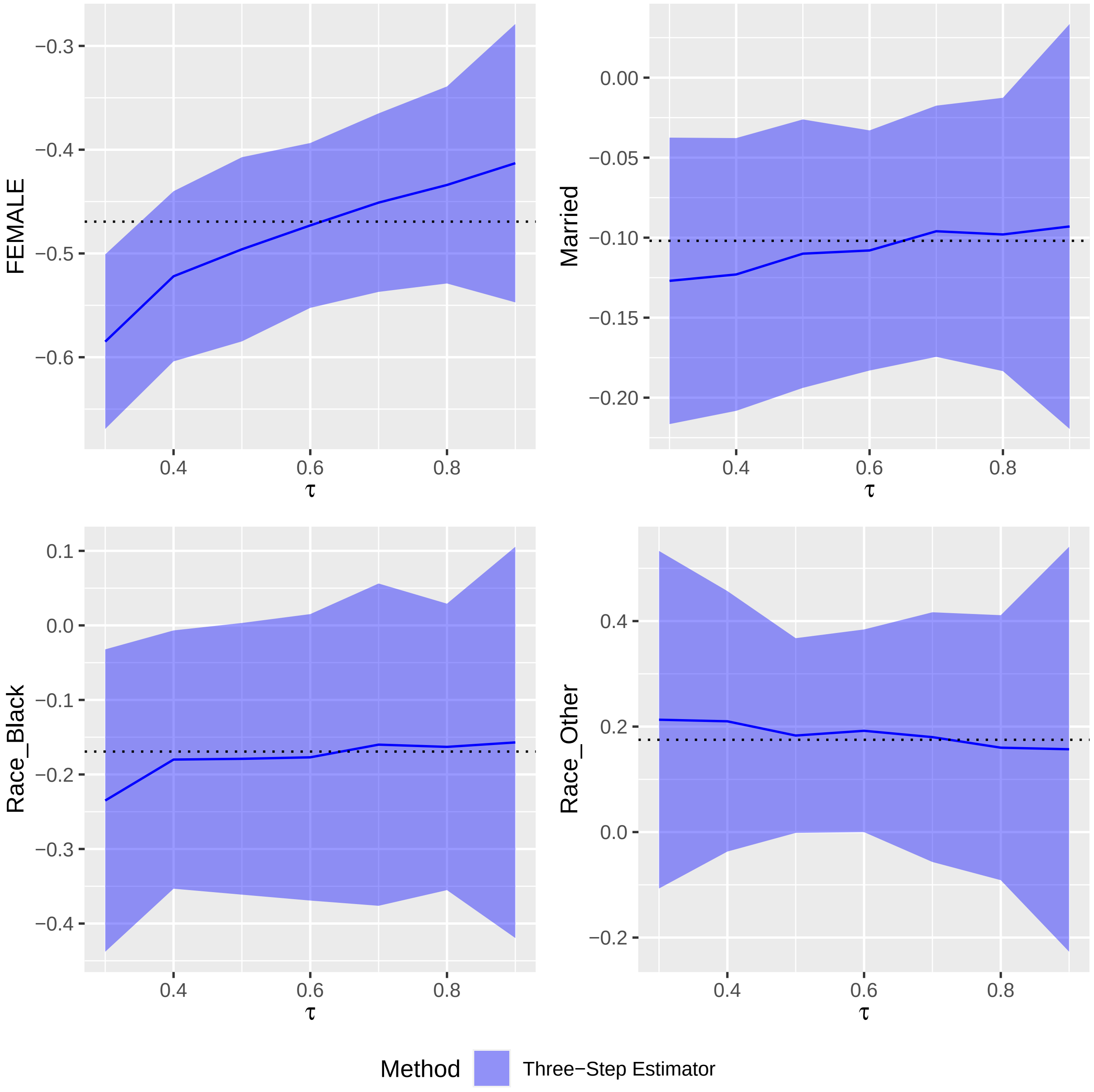


Figure 4.3: Estimated regression coefficients, $\hat{\boldsymbol{\beta}}^{\tau}$, for demographic variables. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

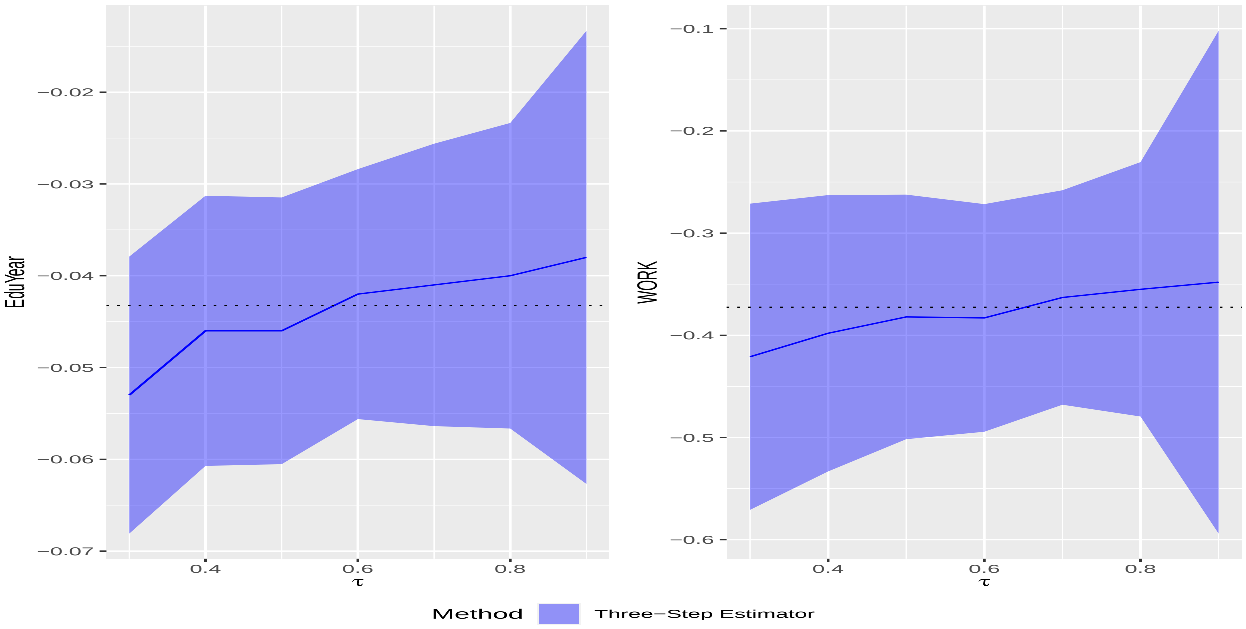


Figure 4.4: Estimated quantile regression coefficients, $\hat{\boldsymbol{\beta}}^{\tau}$, for socio-economic variables. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

final model. Our analysis found similar estimates between the mean structure and the quantile functions, except for the highly significant predictors. This indicates the effects do not vary across the conditional distribution, which is not commonly observed in quantile regression analyses; on the other hand, this pattern could help inform more advanced experimental design for similar future studies.

Based on our analysis, prevention of heavy drinking is important as the effects of policies and treatments are stronger. When patients are not heavy alcoholics and do not drink too much, relatively speaking, the medication and behavioral intervention tend to work better; however, if patients become heavily alcohol dependent and cannot control their alcohol consumption, the efficacy of therapy diminishes. For the best outcomes, our model suggests the combination of medication and behavioral care at the same time.

4.6 Appendix

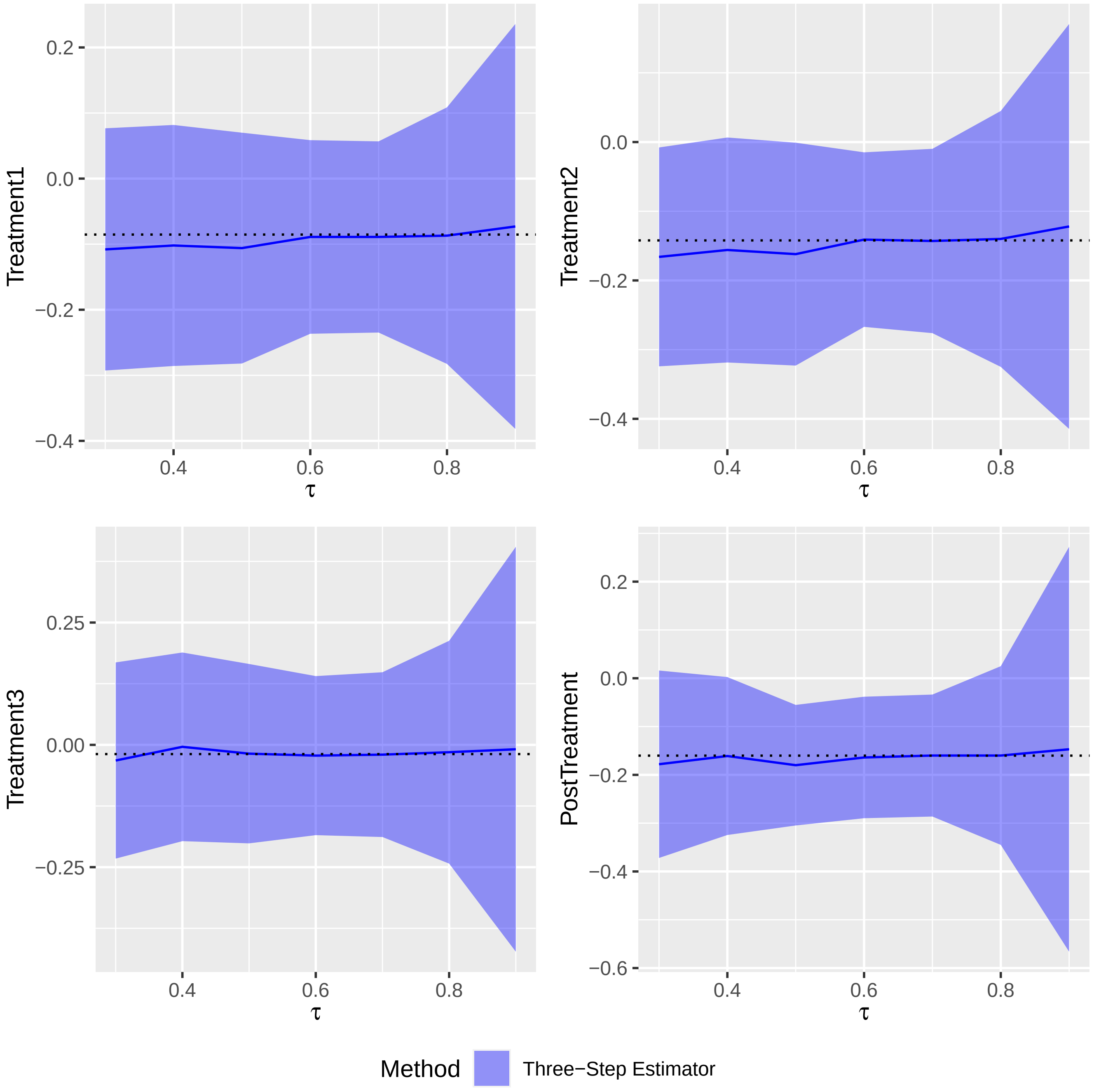


Figure 4.5: Estimated regression coefficients, $\hat{\boldsymbol{\beta}}^{\tau}$, for therapy combination variables. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

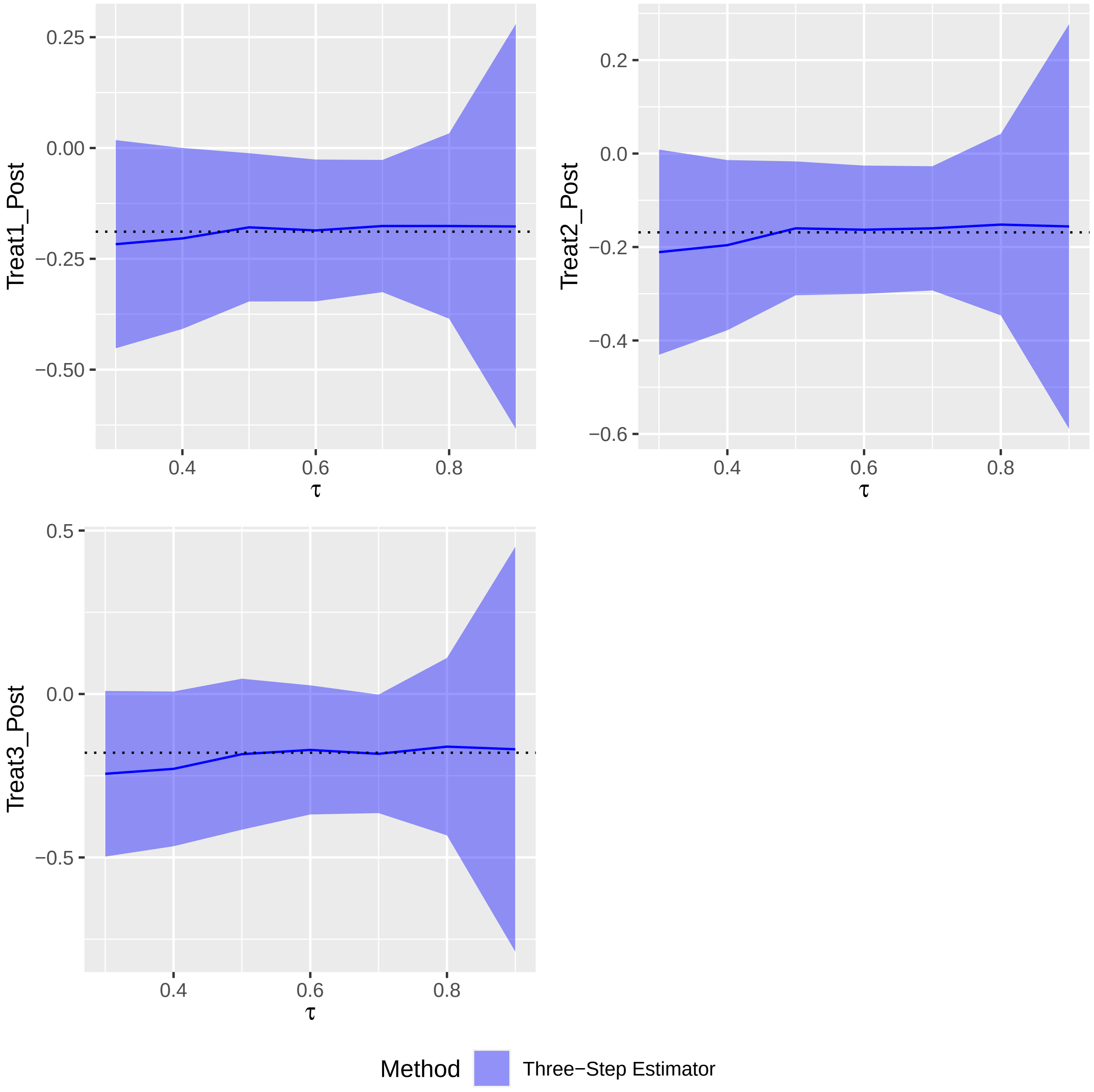


Figure 4.6: Estimated regression coefficients, $\hat{\boldsymbol{\beta}}^{\tau}$, for interaction between treatments and treatment periods. The dotted line is the estimated value of β for the mean structure, obtained by ZI GLMM model in the first step with ZINB specification

Table 4.5: Variable definitions and summary statistics for the COMBINE Study dataset.

Variables	Definition	Mean	Min	25% Quantile	Median	75% Quantile	Max
MEAN	Average number of drinks per day	3.393	0	0	2	5	49
FEMALE	If subject is female: 1	0.304	0	0	0	1	1
MARRIED	Indicator for marital status	0.468	0	0	0	1	1
RACE(Black)	If subject is black or African American:1	0.072	0	0	0	0	1
RACE(Other)	If Race(Other): 1	0.039	0	0	0	0	1
EduYear	Years of education completed	14.600	2.000	12.000	14.000	16.000	30.000
Treatment1	If subject received medication only: 1	0.330	0	0	0	1	1
Treatment2	If subject received medication plus CBI: 1	0.450	0	0	0	1	1
Treatment3	If subject received CBI only: 1	0.105	0	0	0	0	1
Work	Indicator for current employment status	0.904	0	1	1	1	1
PostTreatment	If measurement taken after treatment (week 26, 52, 68): 1	0.475	0	0	0	1	1

Chapter 5 Discussion and Future Study

5.1 Discussion

Chapter 1 of this dissertation provided a comprehensive review of count regression models and QR model. In particular, we focused on the ZI regression model as our research's main focus was on data with zero-inflation. We then discussed the literature for the ZI regression model in longitudinal settings, while the current literature on QR for count/ZI data is almost minimal.

Chapter 2 formally introduced our three-step model for characterizing the conditional quantiles of count data and ZI data. The first step models the count responses' conditional mean structure, where all existing models can be incorporated into our framework. The class of GLMMs and ZI GLMMs are effective classes of models for this step, based on parametric model fitting. Two main distributions, Poisson and negative binomial, are good choices for the count data. Their extensions to the ZI settings, ZIP model and ZINB model, provide reasonable estimates under the circumstance of zero-inflation. The second step utilizes interpolation to make the transition from the mean structure to the quantile functions. This step leverages the connection between a discrete distribution and its continuous counterpart. Finally, the third step regresses the quantiles on the independent variables. This provides a novel estimation routine to investigate the effects of predictors on the dependent variable's conditional quantiles. Inferential considerations based on presumed asymptotic theory as well as bootstrap methods were given to compliment the estimation procedure. Non-parametric bootstrap, such as the multiplier bootstrap (also known as weighted bootstrap), is straightforward. Methods for model selection and GOF assessment are also provided in this chapter. Extensive simulations were conducted to check the performance of the proposed method. An empirical application to the OHIE data was provided to illustrate the application to real data.

In Chapter 3, we extended our method to longitudinal settings, where repeated measurements were recorded for the same unit. We provided computational details of such an extension, emphasizing subject heterogeneity while accounting for zero-inflation and over-dispersion. An extensive simulation study also supported our method's performance. We analyzed data from the famous RAND Health Insurance Experiment to provide new insights for the modeling strategies.

Chapter 4 discussed another empirical analysis for alcohol dependency, which is of great practical importance. By applying our method to the COMBINE Study data, we provided a novel and comprehensive characterization for policy-making and public health. Given the economic and social costs caused by heavy drinking, our analysis is informative for the disease of alcoholism.

In summary, we proposed a three-step approach for modeling the conditional quantiles and measuring the effects of predictor variables on multiple quantiles' responses. This novel method works for regular count data and extends to ZI data and longitudinal settings. Our approach is parametric in the first step, where mean regression models such as the Poisson regression model and ZIP regression model can be applied directly. Computationally, we found modeling via the `R` package `glmmTMB` (Brooks et al., 2017) to be convenient and powerful since the most popular choices of distributions were already implemented in this package.

Parametric models are usually less robust compared to semi-parametric models or non-parametric models. Hence, it is worth noting that semi-parametric or non-parametric methods can carry out our approach's first step. On the other hand, model diagnostics and validation are important for our approach. They help check the parametric structures, determine the best model fit, and improve the estimation performance by decreasing the bias and variance. As the simulation results show, a model's performance is excellent when a good model is fitted to the data. Hence, we emphasized the importance of model checking via different techniques, such as information criterion and residuals analysis, and incorporated them as an essential part of our modeling strategy. Based on our study, information criteria, such as AIC or BIC, are useful for variable selection and model comparison. Boundary LR tests can be used to test for the presence of zero-inflation and over-dispersion, thus helping researchers determine which model to use. This is very useful when, for example, a comparison between a negative binomial regression and a ZINB regression is to be made.

Another model checking tool is the randomized quantile residuals (Dunn and Smyth, 1996), which provides a straightforward visualization for illustration.

Our approach then uses interpolation of the discrete distribution for the count model under consideration. The resulting distribution is a continuous counterpart of the selected discrete distribution. Certain important features are the same between the two versions. Firstly, the canonical parameters are the same for both the discrete version and the continuous version. For example, the mean (and variance) parameter λ from a Poisson distribution stays the same when interpolation introduces the corresponding continuous Poisson distribution. This is the reason why the estimation in the first step is useful, since the estimated parameters are always the same for the different versions of distribution. Secondly, the quantile functions for both versions match at integer values. Given that the observed values are always counts, this feature sets up the transition to the quantile estimation in the third step.

The final step uses the NLS estimation for the effects of predictors on the quantiles of the response. The NLS method provides a flexible routine to estimate the specified functional forms. For the longitudinal settings, an extension of NLS estimation that includes random effects is utilized. In our application, we consider individual intercepts for subject heterogeneity, but NLS can also handle more complicated model specifications. Another point worth noting is that the NLS method requires the specification of a functional form. For our study, an exponential function is used as it is the assumed form in the QR literature. This assumption can be relaxed if a non-parametric routine is used in the third step, and this can be a potential extension for future study.

The aforementioned framework works for cross-sectional data and longitudinal data. The efficacy of our procedure is highlighted by applications to the Oregon Health Insurance data, the RAND Health Insurance Experiment data and the COMBINE Study. By capturing the specific patterns of the data, our analyses provide better fits and yield more accurate characterization.

5.2 Future Study

Given the limitations of our study, we proposed the following directions for research study. These points are of great importance and can be interesting topics on their own.

The first point we have is to derive more comprehensive asymptotic results regarding the three-step estimator. This is important for researchers to better understand the properties of three-step estimator, and efficient for inferences. In Chapter 2, we derived some asymptotic results for the sampling distribution of the estimator; however, given the complexity of the problem, we turned to bootstrap for the inferences, including the estimation of SEs and the construction of confidence intervals. Bootstrap methods, albeit straightforward and easy to implement, are computationally expensive. If researchers need to analyze a large dataset, a good direction is to employ asymptotic theories. In their separate literature, asymptotic results under certain circumstances have been derived for the methods in the first step (Gan, 2000) and the third step (Wu, 1981; Kundu, 1993). These resources provide a good foundation for the asymptotic theories of our models.

The other potential direction for future study is to relax the parametric assumption by introducing more robust estimation. This can be accomplished by using semi-parametric or non-parametric estimations in the first step and the third step. Given the flexibility of our strategies, this extension is actually natural and easy to implement. The main obstacle, however, is in the second step. When no assumption is made on the distribution, the connection between the mean structure and the quantile structure is unclear. To overcome this difficulty, one potential direction is to combine the first two steps into one integral step, where the conditional quantiles can be estimated directly.

Bibliography

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

Anton, R. F., S. S. O'Malley, D. A. Ciraulo, R. A. Cisler, D. Couper, D. M. Donovan, D. R. Gastfriend, J. D. Hosking, B. A. Johnson, J. S. LoCastro, R. Longabaugh, B. J. Mason, M. E. Mattson, W. R. Miller, H. M. Pettinati, C. L. Randall, R. Swift, R. D. Weiss, L. D. Williams, A. Zweben, and for the COMBINE Study Research Group (2006, 05). Combined Pharmacotherapies and Behavioral Interventions for Alcohol DependenceThe COMBINE Study: A Randomized Controlled Trial. *JAMA 295*(17), 2003–2017.

Baetschmann, G. and R. Winkelmann (2012). Modelling Zero-Inflated Count Data when Exposure Varies: With an Application to Sick Leave. Technical Report 61, Department of Economics, University of Zurich.

Baicker, K., A. Finkelstein, J. Song, and S. Taubman (2014). The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment. *American Economic Review 104*(5), 322–328.

Baicker, K., S. Taubman, H. Allen, M. Bernstein, J. Gruber, J. Newhouse, E. Schneider, B. Wright, A. Zaslavsky, and A. Finkelstein (2013). The Oregon Experiment - Effects of Medicaid on Clinical Outcomes. *New England Journal of Medicine 368*(18), 1713–1722.

Barrodale, I. and F. Roberts (1978). An efficient algorithm for discrete l1 linear approximation with linear constraints. *SIAM Journal on Numerical Analysis 10*(5), 603–611.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software 67*(1), 1–48.

Bohning, D., E. Dietz, P. Schlattmann, L. Mendonca, and U. Kirchner (1999). The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental

Epidemiology. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 162*(2), 195–209.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Mächler, and B. M. Bolker (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal 9*(2), 378–400.

Cameron, A. C. and P. K. Trivedi (2013). *Regression Analysis of Count Data* (2nd ed.). Cambridge, UK: Cambridge University Press.

Chaudhuri, P. (1991). Nonparametric Esitmates of Regression Quantiles and Their Local Bahadur Representation. *The Annals of Statistics 19*(2), 760–777.

Chernozhukov, V., I. Fernandez-Val, B. Melly, and K. Wuthrich (2018). Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes. *Preprint*.

Chernozhukov, V., I. Fernández-Val, B. Melly, and K. Wüthrich (2020). Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes. *Journal of the American Statistical Association 115*(529), 123–137.

Chernozhukov, V., I. Fernández-Val, and M. Weidner (2017). Network and Panel Quantile Effects Via Distribution Regression. mimeo.

Chernozhukov, V., I. Fernandez-Val, and M. Weidner (2017). Network and Panel Quantile Effects Via Distribution Regression. *mimeo*.

Deb, P. and P. Trivedi (2002). The Structure of Demand for Health Care: Latent Class Versus Two-Part Models. *Journal of Health Economics 21*(4), 601–625.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 39*(1), 1–38.

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

Edwards, G. and M. M. Gross (1976). Alcohol dependence: provisional description of a clinical syndrome. *British medical journal 1*(6017), 1058–1061.

Efron, B. (1992). Poisson Overdispersion Estimates Based on the Method of Asymmetric

Maximum Likelihood. *Journal of the American Statistical Association 87*, 98–107.

Fang, R., B. D. Wagner, J. K. Harris, and S. A. Fillon (2016). Zero-Inflated Negative Binomial Mixed Model: An Application to Two Microbial Organisms Important in Oesophagitis. *Epidemiology and Infection 144*(11), 2447–2455.

Feng, J. and Z. Zhu (2011). Semiparametric Analysis of Longitudinal Zero-Inflated Count Data. *Journal of Multivariate Analysis 102*(1), 61–72.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. Newhouse, H. Allen, K. Baicker, and O. Group (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. *The Quarterly Journal of Economics 127*(3), 1057–1106.

Galton, F. (1883). *Inquiries into human faculty and its development*. JM Dent and Company.

Gan, N. (2000). *Generalized Zero-Inflated Models and Their Applications*. Ph. D. thesis, North Carolina State University.

Geraci, M. (2014). Linear Quantile Mixed Models: The lqmm Package for Laplace Quantile Regression. *Journal of Statistical Software 57*(13), 1–29.

Geraci, M. (2016). Qtools: A Collection of Models and Tools for Quantile Inference. *The R Journal 8*(2), 117–138.

Geraci, M. and M. Bottai (2007). Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution. *Biostatistics 8*(1), 140–154.

Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. *Working Paper EC-94-10: Department of Economics, New York University. SSRN 1293115*.

Greenfield, S. F., H. M. Pettinati, S. O'Malley, P. K. Randall, and C. L. Randall (2010). Gender Differences in Alcohol Treatment: An Analysis of Outcome From the COMBINE Study. *Alcohol Clin Exp Res 34*(10), 1803–1812.

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

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

Harding, M. and C. Lamarche (2019a). Penalized Estimation of a Quantile Count Model for Panel Data. *Annals of Economics and Statistics* (134), 177–206.

Harding, M. and C. Lamarche (2019b). Penalized Estimation of a Quantile Count Model for Panel Data. *Annals of Economics and Statistics* (134), 177–206.

Hilbe, J. M. (2011). *Negative Binomial Regression* (2^{nd} ed.). Cambridge, UK: Cambridge University Press.

Iddi, S. and G. Molenberghs (2013). A Marginalized Model for Zero-Inflated, Overdispersed and Correlated Count Data. *Electronic Journal of Applied Statistical Analysis 6*(2), 149–165.

Ilienko, A. (2013a). Continuous counterparts of Poisson and binomial distributions and their properties. *Preprint*.

Ilienko, A. (2013b). Continuous Counterparts of Poisson and Binomial Distributions and Their Properties. arXiv:1303.5990 [math.PR].

Jansakul, N. and J. Hinde (2002). Score Tests for Zero-Inflated Poisson Models. *Computational Statistics and Data Analysis 40*(1), 75–96.

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

Jørgensen, B. (1987). Exponential Dispersion Models (with Discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 49*(2), 127–162.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

Koenker, R. (2017). Quantile Regression: 40 Years On. *Annual Review of Economics 9*(1), 155–176.

Koenker, R. and G. Bassett (1978). Regression Quantiles. *Econometrica 46*(1), 33–50.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). *Handbook of Quantile Regression*. CRC Press.

Koenker, R. and B. Park (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics 71*(1-2), 265–283.

Kundu, D. (1993). Asymptotic theory of least squares estimator of a particular nonlinear regression model. *Statistics and Probability Letters 18*(1), 13–17.

Lamarche, C. (2010). Robust penalized quantile regression estimation for panel data. *Journal of Econometrics 157*, 396–408.

Lamarche, C., X. Shi, and D. S. Young (2021). Conditional Quantile Functions for Zero-Inflated Longitudinal Count Data. submitted.

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

Lee, D. and T. Neocleous (2010). Bayesian Quantile Regression for Count Data with Application to Environmental Epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(5), 905–920.

Li, C. (2012). Identifiability of Zero-Inflated Poisson Models. *Brazilian Journal of Probability and Statistics 26*(3), 306–312.

Liang, K.-Y., S. L. Zeger, and B. Qaqish (1992). Multivariate Regression Analyses for Categorical Data. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 54*(1), 3–40.

Lindstrom, M. J. and D. M. Bates (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics 46*(3), 673–687.

Loh, W.-Y. (1991). Bootstrap Calibration for Confidence Interval Construction and Selection. *Statistica Sinica 1*(2), 477–491.

Lord, D., S. Washington, and J. Ivan (2005). Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis and Prevention 37*(1), 35–46.

Ma, S. and M. Kosorok (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis 96*, 190–217.

Machado, J. and J. Santos Silva (2005). Quantiles for Counts. *Journal of the American Statistical Association 100*(472), 1226–1237.

Mwalili, S., E. Lesaffre, and D. Declerck (2008). The Zero-Inflated Negative Binomial

Regression Model With Correction for Misclassification: An Example in Caries Research. *Statistical Methods in Medical Research 17*(2), 123–139.

Newey, W. and J. Powell (1987). Asymmetric Least Squares Estimation and Testing. *Econometrica 55*, 819–847.

Padellini, T. and H. Rue (2019a). Model-Aware Quantile Regression for Discrete Data. arXiv:1804.03714v2 [stat.ME].

Padellini, T. and H. Rue (2019b). Model-Aware Quantile Regression for Discrete Data. *Preprint*.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-Plus*. New York, NY: Springer.

Portnoy, S. and R. Koenker (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science 12*(4), 279–300.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ridout, M., J. Hinde, and C. Demetrio (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics 57*(1), 219–223.

Sawa, T. (1978). Information Criteria for Discriminating Among Alternative Regression Models. *Econometrica 46*(6), 1273–1291.

Stroup, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Boca Raton, FL: CRC Press.

Takeuchi, I., Q. Le, T. Sears, and A. Smola (2006). Nonparametric Quantile Regression. *Journal of Machine Learning Research 7*, 1231–1264.

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

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

Wang, K., K. K. W. Yau, and A. H. Lee (2002). 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.

Wang, W., X. Wu, X. Zhao, and X. Zhou (2020). Quantile Regression for Panel Count Data Based on Quadratic Inference Functions. *Journal of Statistical Planning and Inference 207*, 230 – 245.

Williams, O. and J. Grizzle (1972). Analysis of contingency tables having ordered response categories. *Journal of the American Statistical Association 67*(337), 55–63.

Wu, C.-F. (1981). Asymptotic Theory of Nonlinear Least Squares Estimation. *The Annals of Statistics 9*(3), 501 – 513.

Yau, K. K. W., A. H. Lee, and P. J. W. Carrivick (2004). Modeling Zero-Inflated Count Series with Application to Occupational Health. *Computer Methods and Programs in Biomedicine 74*(1), 47–52.

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

Young, D. S., A. M. Raim, and N. R. Johnson (2017). 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.

Young, D. S., E. S. Roemmele, and X. Shi (2021). Zero-Inflated Modeling Part II: Zero-Inflated Models for Complex Data Structures. *WIREs Computational Statistics (in press)*.

Young, D. S., E. S. Roemmele, and P. Yeh (2021). Zero-Inflated Modeling Part I: Traditional Zero-Inflated Count Regression Models, Their Applications, and Computational Tools. *WIREs Computational Statistics (in press)*.

Yu, K., Z. Lu, and J. Stander (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(3), 331–350.

Yu, K. and R. Moyeed (2001). Bayesian quantile regression. *Statistics and Probability*

Letters 54(4), 437–447.

Yue, Y. and H. Rue (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics and Data Analysis 55*(1), 84–96.

Zhu, F. (2012). Zero-Inflated Poisson and Negative Binomial Integer-Valued GARCH Models. *Journal of Statistical Planning and Inference 142*(4), 826–839.

Zweben, A., H. M. Pettinati, R. D. Weiss, M. Youngblood, C. E. Cox, M. E. Mattson, P. Gorroochurn, and D. Ciraulo (2008). Relationship Between Medication Adherence and Treatment Outcomes: the COMBINE Study. *Alcohol Clin Exp Res 32*(9), 1661–9.

Vita

Xuan Shi

EDUCATION

Institution	Major	Degree	Year
University of Kentucky	Statistics	Ph.D.	2021(expected)
University of Kentucky	Statistics	M.S.	2018
Ohio State University	Applied Statistics	M.S.	2015
Central China Normal University	Statistics and Business	B.S.	2013

WORKING EXPERIENCE

- 2019–2019: **Research Grant Statistician**,
  Department of Statistics, University of Kentucky, Lexington, KY
- 2018–2020: **Graduate Teaching Assistant (Primary Instructor)**,
  Department of Statistics, University of Kentucky, Lexington, KY
- 2017–2018: **Graduate Research Assistant**,
  Department of Statistics, University of Kentucky, Lexington, KY
- 2016–2021: **Graduate Teaching Assistant**,
  Department of Statistics, University of Kentucky, Lexington, KY

PUBLICATIONS

- **X. Shi**, D. S. Young, and C. E. Lamarche (2021). “Modeling Strategies for Quantile Regression with Zero-Inflated Discrete Responses.” *(In preparation)*.
- C. E. Lamarche, **X. Shi**, and D. S. Young (2021). “Conditional Quantile Functions for Zero-Inflated Longitudinal Count Data.” Major revision requested.
- D. S. Young, E. S. Roemmele, and **X. Shi**, (2021). “Zero-Inflated Modeling Part II: Zero-Inflated Models for Complex Data Structures.” *WIREs Computational Statistics (in press).*