Conditional Quantile Functions for Zero-Inflated Longitudinal Count Data

Carlos Lamarche*,[a], Xuan Shi[b], Derek S. Young[b]

[a]*Department of Economics, University of Kentucky, Lexington, KY, 40506, USA*
[b]*Dr. Bing Zhang Department of Statistics, University of Kentucky, Lexington, KY, 40536, USA*

Abstract

The identification and estimation of conditional quantile functions for count responses using longitudinal data are considered. The approach is based on a continuous approximation to distribution functions for count responses within a class of parametric models that are commonly employed. It is first shown that conditional quantile functions for count responses are identified in zero-inflated models with subject heterogeneity. Then, a simple three-step approach is developed to estimate the effects of covariates on the quantiles of the response variable. A simulation study is presented to show the small sample performance of the estimator. Finally, the advantages of the proposed estimator in relation to some existing methods is illustrated by estimating a model of annual visits to physicians using data from a health insurance experiment.

Key words: Zero-inflated count data, Quantile models, Subject heterogeneity, Generalized linear mixed models

1. Introduction

Quantile regression, as originally introduced by Koenker and Bassett (1978), is a widely-used approach to estimate flexible models in economics and statistics. When the response is continuous, quantile regression allows practitioners to estimate conditional quantile functions. While theoretical and methodological research over the last 40 years have addressed 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 distribu-

*Corresponding author. 223G Gatton College of Business and Economics, Lexington, KY, USA; Phone: 1-859-257-3371.

Email addresses: clamarche@uky.edu (Carlos Lamarche), shi.517@uky.edu (Xuan Shi), derek.young@uky.edu (Derek S. Young)

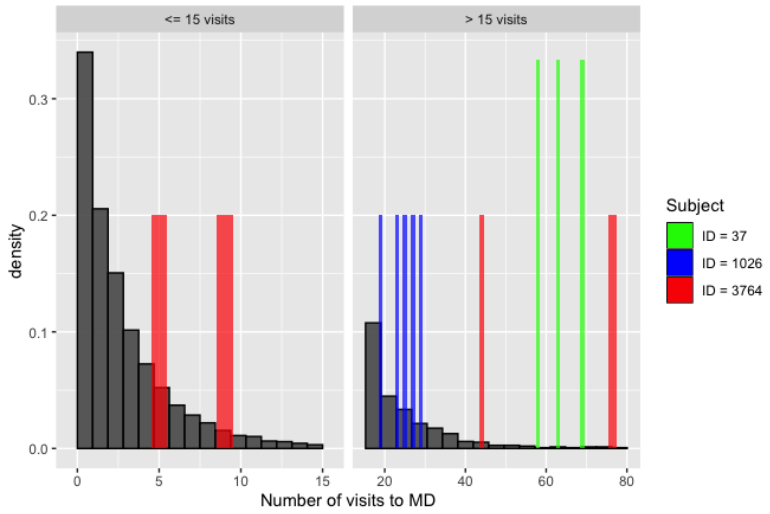

Figure 1: *Number of visits to physicians in the RAND Health Insurance Experiment data. The distribution of the number of visits for three subjects whose observed counts were mostly in the tail of this distribution (the right panel) are overlaid.*

tion of the count variable, but policy recommendations can only be based on average effects.

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. An illustrative example includes the number of visits to physicians and the demand for medical services. Using the RAND Health Insurance Experiment data (Deb and Trivedi, 2002), Figure 1 shows that the proportion of zero visits to physicians exceeds 30% for patients with no greater than 15 visits per year. Moreover, the distribution of the count response has a long tail reaching a maximum of 77 visits, while the average is 2.86. As discussed in Section 5, over 90% of the participants in this experiment are observed over either 3 or 5 years. The distributions of the number of visits for three subjects are also highlighted in Figure 1. These subjects were highlighted as they noticeably contributed to the tail of this distribution, thus suggesting the appropriateness of reflecting latent subject heterogeneity in any proposed model for these data. As in many other applications, it is immediately apparent the need for a flexible approach that simultaneously addresses zero inflation and latent subject heterogeneity, while allowing estimation of the effects of covariates across the conditional distribution of the count response variable.

In order to address these challenges, this paper investigates estimation of conditional quantile functions and covariate effects for longitudinal count responses. Our approach is based on a continuous approximation to distribution functions for count data within a class of parametric models commonly employed in the literature. We adopt an approach based on interpolation of functions for count responses as in Ilienko (2013) and Padellini and Rue (2019), which can be viewed as an alternative smoothing method to the jittering approach proposed by Machado and Santos Silva (2005) and adopted by Harding and Lamarche (2019). We develop a three-step estimation procedure using a generalized linear mixed model (GLMM) procedure, which provides a flexible statistical framework to handle over/underdispersion, shrinkage estimation, and smoothing of regression relationships. In the first step, we consider 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 the third step, a flexible GLMM is employed for a model of conditional quantile responses. The finite sample performance of the estimator is investigated using a simulation study, and we find that the estimator has satisfactory performance for the estimation of quantile effects under different degrees of zero inflation.

Our work 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 (2021) developed an approach based on distribution regression. The literature on panel quantiles includes just a few papers. Harding and Lamarche (2019) 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, including distributional regression (Kneib et al., 2021), 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 (2019). 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, 2019).

As highlighted in the preceding discussion, one of the contributions of this work is how we address zero inflation in longitudinal data. Zero inflation occurs when zero counts arise from one of two possible states: a degenerate state or from a 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 ZI Poisson (ZIP) regression model as a way to characterize zero defects in a manufacturing process that manifest from one of two states: a *perfect state* where defects are extremely rare or an *imperfect*

state where defects are possible. Since then, numerous extensions to the ZIP regression model have been developed; see Young et al. (2021b) and Young et al. (2021a) for a contemporary review, and Cameron and Trivedi (2013) for a summary of econometric analysis with count data. In particular, just like in non-ZI models, random effects have been included in ZI models to capture various features of the data, such as subject heterogeneity (Zhu et al., 2017), 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 use random individual intercepts to account for subject heterogeneity when estimating conditional quantiles for longitudinal data with ZI count responses.

This paper is organized as follows. In Section 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, 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. We note that our focus is strictly on ZI GLMMs with count responses and excludes the setting of semi-continuous responses, such as the ZI gamma and ZI lognormal distributions. In Section 4, we provide an extensive simulation study to assess the performance of our approach in estimating mean and quantile effects. In Section 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 6.

2. Conditional Quantiles of Count Responses

Suppose that we randomly sample N subjects where the i^{th} unit 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} measurement on the i^{th} unit is a vector of p observed independent variables, given by $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itp})^\top$. We also define $\mathbf{w}_{it} = (w_{it1}, \ldots, w_{itd})^\top$ to be a vector of d observed independent variables, which will be used to model the 0 counts in our models discussed below. Note that $\mathbf{w}_{it}$ can simply be $\mathbf{x}_{it}$, it may contain a subset of the variables in $\mathbf{x}_{it}$ along with some independent variables not included in $\mathbf{x}_{it}$, or it may be an entirely different set of independent variables from those in $\mathbf{x}_{it}$. Assume further that with the t^{th} measurement on the i^{th} unit is a vector of q design variables for the random effects, given by $\mathbf{z}_{it} = (z_{it1}, \ldots, z_{itq})^\top$. The $\mathbf{z}_{it}$ have a specific structure (typically) composed of 0s and 1s to reflect the subject-specific quantities in the model. For example, if interested in subject-specific intercepts (a random effect), then $\mathbf{z}_{it}$ will just be univariate and simply equal to 1.

Let $\theta_{it} = \mathrm{E}[y_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}]$ denote the conditional mean of the parametric distribution $F_{y_{it}}$ of the count response, y_{it}. Specifically, $F_{y_{it}}$ is the cumulative

distribution for a parametric discrete distribution, such as the Poisson or negative binomial. Let $G_{y_{it}}$ be the cumulative distribution function of a ZI count variable,

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

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

We propose to consider the following continuous counterpart to the ZI count distribution (1):

$$G_{y'_{it}}(y) = \pi_{it} + (1 - \pi_{it})k(y, \theta_{it}), \tag{2}$$

where $k(y, \theta_{it}) = F_{y'_{it}}(y)$ is the cumulative distribution function 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 Ilienko (2013) and Padellini and Rue (2019) for similar derivations in models without zero inflation.

The approximation (2) can be used on the two leading distributions for count models: ZIP and zero-inflated negative binomial (ZINB). 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 that, for $y > -1$,

$$G_{y'_{it}}(y) = \pi_{it} + (1 - \pi_{it})\frac{\Gamma(y + 1, \theta_{it})}{\Gamma(y + 1)}. \tag{3}$$

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

$$G_{y'_{it}}(y) = \pi_{it} + (1 - \pi_{it})I_{1-p_{it}}(r, y + 1) = \pi_{it} + (1 - \pi_{it})\frac{B(r, y + 1, 1 - p_{it})}{B(r, y + 1)}, \tag{4}$$

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. This leads us to our first result.

Proposition 1. *The continuous zero-inflated distributions functions* (3) *and* (4) *are well-defined distribution functions.*

Using the previous expressions, we have a recursive model and the parameters of the model are identified. For instance, in the case of a ZIP model with a logit link generalized linear model used for the zero-inflation probability, we have,

$$\pi_{it} = \frac{\exp(\mathbf{w}_{it}^\top \boldsymbol{\gamma})}{(1 + \exp(\mathbf{w}_{it}^\top \boldsymbol{\gamma}))},$$

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

and

$$\tau = \pi_{it} + (1 - \pi_{it}) \frac{\Gamma(y + 1, \theta_{it})}{\Gamma(y + 1)},$$

where $\mathbf{w}_{it}$ is a vector of independent variables, possibly different than $\mathbf{x}_{it}$, and $\tau \in (0, 1)$ is a quantile of the distribution of the continuous approximation to the discrete distribution for all natural numbers y. 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, y', in a model with mean θ_{it} and probability π_{it}.

Let $K_{it}(y, \tau) = (\tau - \pi_{it}) - (1 - \pi_{it}) k(y, \theta_{it})$. While identification of π_{it} and θ_{it} is standard in the literature (see, for example, Li, 2012), the next result shows how to obtain conditional quantile values of the response variable.

Proposition 2. *Let $\tau \in (0, 1)$ and $k(y, \theta_{it})$ be a continuous increasing variable in its first argument. For $\tau > \pi_{it}$ for all i, t, the solution to $K_{it}(y, \tau) = 0$, say, y_{it}^τ, exists and it is unique.*

Having identification of the triple $(\theta_{it}, \pi_{it}, y_{it}^\tau)$, we obtain the quantile-specific effects on the count response variable as the argument that minimizes the expected loss,

$$\mathrm{E}\{L(y_{it}^\tau - h(\eta_{it}))\}, \tag{5}$$

where $L(\cdot)$ is a loss function, $h(\cdot)$ is an inverse link function, and η_{it} is a (linear) predictor variable which is determined by regressors and individual intercepts. Specifically, the η_{it} are implicitly a function of the $\boldsymbol{\beta}$ and $\mathbf{u}_i$ presented in the above discussion. While (5) offers a general formulation of the problem that can be accommodated for other models, the next section offers specific forms for the loss function, link function, and the predictor variables employed in the empirical sections.

3. Model Specification and Estimation

Let the $\mathbf{x}_{it}^\top$ form the rows of the $T \times p$ design matrix $\mathbf{X}_i$. Moreover, let the $\mathbf{z}_{it}^\top$ form the rows of the $T \times q$ design matrix $\mathbf{Z}_i$, which can be, for example, simple basis functions of a time index. Both $\mathbf{X}_i$ and $\mathbf{Z}_i$ typically include a column of 1s to permit estimation of, respectively, an overall intercept and a subject-level intercept.

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

$$\boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i, \tag{6}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\mathbf{u}_i \in \mathbb{R}^q$ are, respectively, fixed-effects coefficient and random-effects coefficient vectors in the mixed models literature. Letting $\boldsymbol{\mu}_i | \mathbf{u}_i$ denote the mean of the (conditional) distribution of $\mathbf{y}_i | \mathbf{u}_i$, 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

$h(\cdot) := g^{-1}(\cdot)$ is used to write the inverse link function. Typically, the preceding setup assumes that the $\mathbf{u}_i$ are independent and identically distributed (iid) $\mathcal{N}_q(\mathbf{0}, \mathbf{G})$, where $\mathbf{G}$ is positive definite. Such an assumption is often done for mathematical and computational convenience, but typically performs well in practice. Alternatively, Zhang and Davidian (2001) thoroughly treat relaxing the normality assumption with emphasis on the "semi-nonparametric" representation of Gallant and Nychka (1987), followed by the effective use of information criteria to choose the final distribution on the random effects. Regardless, to further solidify the form of the random component of (6), especially with respect to the application presented in Section 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 is a scalar. Therefore, $q = 1$. However, framing our discussion using the more general form of $\mathbf{Z}_i$ and $\mathbf{u}_i$ in the GLMM linear predictor reflects that our proposed methodology has the capacity to handle more complex mixed-effects structures. Notably, the development of conditional quantile functions for count responses in the GLMM framework differs from existing quantile methods that have been developed in the linear mixed model (LMM) framework (Geraci and Bottai, 2014; Battagliola et al., 2021). Moreover, the ability to handle zero-inflation in the count responses is also part of our contribution.

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 blocks of units, such as repeated measures on the same subject or multiple households measured within defined levels of geography; e.g., census blocks. The framework also accommodates intrasubject correlation, a statistical feature leveraged in the modeling of panel data as in the focus of the present work. The present work is also focused on the setting where the response values of the panel data are counts. Thus, the primary distributions studied for $y_{it}|\mathbf{u}_i$, for which we will generically denote the probability mass function (pmf) as $p_{y_{it}|\mathbf{u}_i}$, will be the Poisson and the negative binomial. Specifically, the gamma-Poisson mixture representation is used for the negative binomial (also known as the NB2 model in Hilbe, 2011), which results in a parameterization that also includes the presence of a dispersion parameter, say, ϕ. We strictly use the NB2 model in the present work, so henceforth it will be understood that any discussion about the negative binomial distribution is in terms of the NB2 parameterization. One challenge with maximum likelihood estimation of such GLMMs is that the marginal likelihood involves integration over a (generally) messy product of Gaussian and exponential family likelihoods (or quasi-likelihoods) due to the random effects. Direct maximization is generally not possible, however, integral approximations via Gauss-Hermite quadrature or Laplace approximations typically perform very well. In `R` (R Core Team, 2019), both approximations are options in the `glmer()` function (and `glmer.nb()` function for estimating negative binomial mixed models) 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 present the model details for ZI GLMMs when the count distribution

is either ZIP or ZINB, followed by highlighting some of the relevant estimation details. For $\mathbf{G}$, the variance-covariance matrix of the random effects $\mathbf{u}_i$, let $\text{vech}(\mathbf{G}) \in \Lambda$, where Λ is 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 zeros in our count outcomes are generated from one of two possible processes: a degenerate distribution ("perfect" state) with probability $\pi_{it} \equiv d(\mathbf{w}_{it}^\top \boldsymbol{\gamma})$ or the count distribution $p_{y_{it}|\mathbf{u}_i}$ ("imperfect" state) with probability $1-\pi_{it}$. Therefore,

$$y_{it}|\mathbf{u}_i \sim \begin{cases} 0, & \text{with probability } \pi_{it}; \\ p_{y_{it}|\mathbf{u}_i}, & \text{with probability } 1-\pi_{it}. \end{cases} \tag{7}$$

Here, π_{it} is again a probability that determines how we choose between the two states. Thus, $d^{-1}(\cdot)$ is taken as a logit link function as it is the natural link that linearizes such Bernoulli probabilities of "success." The linearization involves an r-dimensional vector of regressors, $\mathbf{w}_{it}$, and a parameter vector $\boldsymbol{\gamma} \in \Gamma$, where Γ is an open subset of $\mathbb{R}^r$. Note that $w_{it1} \equiv 1$ also to accommodate an intercept. The pmf for the ZI count variable defined in (7) is thus

$$f_{y_{it}|\mathbf{u}_i}(y_{it}; \mathbf{x}_{it}, \mathbf{w}_{it}, \mathbf{z}_{it}, \mathbf{u}_i, \boldsymbol{\vartheta}) = \begin{cases} \pi_{it} + (1-\pi_{it}) p_{y_{it}|\mathbf{u}_i}(0; \boldsymbol{\xi}, \mathbf{G}), & \text{if } y_{it} = 0; \\ (1-\pi_{it}) p_{y_{it}|\mathbf{u}_i}(y_{it}; \boldsymbol{\xi}, \mathbf{G}), & \text{if } y_{it} \in \mathbb{N}^+, \end{cases} \tag{8}$$

where $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$ and $\boldsymbol{\vartheta} = (\boldsymbol{\xi}^\top, \boldsymbol{\gamma}^\top, \text{vech}(\mathbf{G})^\top)^\top \in \Theta = \Xi \times \Gamma \times \Lambda$. The above framework defines the model for a ZI GLMM and our immediate concern is obtaining estimates of this model.

The classic ZI count regression models (e.g., ZIP regression and ZINB regression) define the conditional distribution of the ZI GLMM. The ZI GLMM follows the same setup as the GLMM presentation given around Equation (6), but using the ZI distribution given in (7). Estimation can be performed a number of different ways, each carrying their own set of challenges. A seemingly natural approach is to first consider maximum likelihood estimation. The loglikelihood function for which we need to maximize is as follows:

$$\begin{aligned} \ell(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{w}) = &\sum_{(it):y_{it}=0} \sum \log\left\{\pi_{it} + (1-\pi_{it}) p_{y_{it}|\mathbf{u}_i}(y_{it}; \boldsymbol{\xi}, \mathbf{G})\right\} \\ &+ \sum_{(it):y_{it}>0} \sum \left\{\log(1-\pi_{it}) + \log(p_{y_{it}|\mathbf{u}_i}(y_{it}; \boldsymbol{\xi}, \mathbf{G}))\right\}. \end{aligned} \tag{9}$$

The above is clearly difficult to directly optimize, especially given the presence of the random effects $\mathbf{u}_i$. One way to proceed is to consider augmenting the above observed loglikelihood with V_{it}, which is an indicator variable equal to 1 if observation (it) belongs to the degenerate state and 0 otherwise. This augmentation results in the complete data loglikelihood. Treating both the V_{it} and $\mathbf{u}_i$ as

missing data, optimization of the complete data loglikelihood can thus be done via an expectation-maximization (EM) algorithm (Dempster et al., 1977). This setup affords us the luxury of the complete data loglikelihood separating out into a term involving $\boldsymbol{\gamma}$ and another term involving $\log(p_{y_{it}|\mathbf{u}_i}(y_{it}; \boldsymbol{\xi}, \mathbf{G}))$, however, we still have to deal with the integration over the random effects space. This can be accomplished through, for example, (adaptive) Gauss-Hermite quadrature, which is implemented by the `glmer()` function in `lme4` (Bates et al., 2015).

Another approach is to attempt the direct optimization of the observed log-likelihood in (9). The challenge, again, is how the integration is performed over the random effects. A popular approach is to use the Laplace approximation to calculate the marginal likelihood that results from integrating over the random effects space. Let $q(\mathbf{u}, \boldsymbol{\vartheta}) = -\ell(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{w})$ and

$$\mathcal{L}(\boldsymbol{\vartheta}) = \int_{\mathbb{R}^q} \cdots \int \exp\{-q(\mathbf{u}, \boldsymbol{\vartheta}))\} d\mathbf{u}.$$

The maximum likelihood estimate for $\boldsymbol{\vartheta}$ is then

$$\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta}} \mathcal{L}(\boldsymbol{\vartheta}). \tag{10}$$

Moreover, the minimizer of $q(\mathbf{u}, \boldsymbol{\vartheta})$ with respect to $\mathbf{u}$ is

$$\hat{\mathbf{u}}(\boldsymbol{\vartheta}) = \arg\min_{\mathbf{u}} q(\mathbf{u}, \boldsymbol{\vartheta}).$$

The Hessian of $q(\mathbf{u}, \boldsymbol{\vartheta})$ is then calculated with respect to $\mathbf{u}$ evaluated at $\hat{\mathbf{u}}(\boldsymbol{\vartheta})$:

$$H(\boldsymbol{\vartheta}) = \nabla^2_{\mathbf{u}\mathbf{u}^\top} q(\hat{\mathbf{u}}(\boldsymbol{\vartheta}), \boldsymbol{\vartheta}).$$

Finally, the Laplace approximation for the marginal likelihood is

$$\mathcal{L}^*(\boldsymbol{\vartheta}) = (2\pi)^{q/2} |H(\boldsymbol{\vartheta})|^{-1/2} \exp\{-q(\hat{\mathbf{u}}(\boldsymbol{\vartheta}), \boldsymbol{\vartheta}))\}. \tag{11}$$

The above approximation is quite flexible beyond ZI GLMMs and is implemented in the `TMB` package (Kristensen et al., 2016). 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 best linear unbiased estimators (BLUEs), as well as the best linear unbiased predictors (BLUPs) of the $\mathbf{u}_i$.

3.1. Three-Step Estimator

We now state the three-step procedure for constructing the estimated quantile effects, $\hat{\boldsymbol{\beta}}^\tau$, followed by a detailed discussion of the third step.

1. For the assumed ZI GLMM having pmf of the form in (8), find the maximum likelihood estimate for $\boldsymbol{\vartheta}$, $\hat{\boldsymbol{\vartheta}}$, using $\mathcal{L}^*(\boldsymbol{\vartheta})$, the Laplace approximation for the marginal likelihood as defined in (11).

2. Let $\tau \in (0,1)$ be a quantile of the estimated ZI GLMM based on the maximum likelihood estimate $\hat{\boldsymbol{\vartheta}}$. Appealing to Proposition 2, for each i,t, find $y_{it}^{\tau}|(\mathbf{x}_{it},\mathbf{w}_{it},\mathbf{z}_{it})$, which is the solution to $K_{it}(y,\tau)=0$.
3. Letting $\mathbf{y}_i^{\tau}=(y_{i1}^{\tau},\ldots,y_{iT}^{\tau})^{\top}$, find the quantile-specific effects $\hat{\boldsymbol{\beta}}^{\tau}$ using the risk function
$$\mathrm{E}\{L(\mathbf{y}_i^{\tau}-h(\boldsymbol{\eta}_i^{\tau}))\}. \tag{12}$$
The loss function $L(\cdot)$ can be, for example, a (penalized) L_2-loss, such as the one used later in (17).

In order to obtain the estimated quantile effects $\hat{\boldsymbol{\beta}}^{\tau}$ in Step 3 above, we begin by specifying the following nonlinear mixed model (NLMM):

$$\begin{aligned}\mathbf{y}_i^{\tau} &= h(\boldsymbol{\eta}_i^{\tau})+\boldsymbol{\epsilon}_i\\ \boldsymbol{\eta}_i^{\tau} &= \mathbf{X}_i\boldsymbol{\beta}^{\tau}+\mathbf{Z}_i\mathbf{u}_i^{\tau},\end{aligned} \tag{13}$$

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 maximum likelihood estimation for the ZI GLMM. The random effects are again unobserved quantities, so maximum likelihood estimation is based on the marginal density of the responses $\mathbf{y}^{\tau}$,

$$p(\mathbf{y}^{\tau}|\boldsymbol{\beta}^{\tau},\sigma^{2\tau},\mathbf{G}^{\tau})=\int_{\mathbb{R}^q}\cdots\int p(\mathbf{y}^{\tau}|\mathbf{u}^{\tau},\boldsymbol{\beta}^{\tau},\sigma^{2\tau})p(\mathbf{u}^{\tau}|\mathbf{G}^{\tau})d\mathbf{u}^{\tau}, \tag{14}$$

where $p(\mathbf{y}^{\tau}|\boldsymbol{\beta}^{\tau},\sigma^{2\tau},\mathbf{G}^{\tau})$ is the marginal density of the conditional quantile $\mathbf{y}^{\tau}$ given all of the parameters, $p(\mathbf{y}^{\tau}|\mathbf{u}^{\tau},\boldsymbol{\beta}^{\tau},\sigma^{2\tau})$ is the conditional density of $\mathbf{y}^{\tau}$, given the random effects $\mathbf{u}^{\tau}$, and $p(\mathbf{u}^{\tau}|\mathbf{G}^{\tau})$ is the marginal distribution of $\mathbf{u}^{\tau}$. Note that 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. Rewriting the marginal density in (14) in terms of $\boldsymbol{\Delta}^{\tau}$, the loglikelihood of our NLMM in (13) is, thus,

$$\ell(\boldsymbol{\beta}^{\tau},\sigma^{2\tau},\boldsymbol{\Delta}^{\tau}|\mathbf{y}^{\tau})=\sum_{i=1}^{N}p(\mathbf{y}_i^{\tau}|\boldsymbol{\beta}^{\tau},\sigma^{2\tau},\boldsymbol{\Delta}^{\tau}). \tag{15}$$

Estimation of the above loglikelihood can now 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 an LMM estimation step. See Section B in the Appendix for details on the implementation of the approach. It is important to emphasize that the use of the

Laplace approximation for maximum likelihood estimation of ZI GLMMs, and the alternating algorithm of Lindstrom and Bates (1990) for maximum likelihood estimation of NLMMs, are both necessitated by the presence of random effects. Both approaches provide mechanisms to estimate the random effects, which in-turn are used in the calculation of the maximum likelihood estimates of the model parameters.

As noted in the abstract of their paper, Machado and Santos Silva (2005) state "Given the discreteness of the data, some smoothness must be artificially imposed on the problem." We believe that our paradigm is in the same spirit. Specifically, our three-step estimation procedure leverages the continuous (smooth) approximation to find the quantile according to Proposition 2. The appeal with our approach is that we use a parametric model (which one can show is appropriate based on formal tests and model selection criteria when comparing across multiple candidate models) to develop the conditional quantile estimates. Moreover, our approach is model-aware, which affords us the flexibility to explicitly account for zero-inflation and subject heterogeneity. This is accomplished using ZI GLMMs. However, we have the added complexity of needing to approximate the quantiles, which is common under more complex models like the ZI GLMMs. We further note that this is a commonplace strategy to estimate the quantiles from complex models via simulation, and is addressed in, for example, Breidt (2004). Even though the estimated $\boldsymbol{\beta}^\tau$ are heavily-informed by a numerical procedure, they still retain their interpretation as quantile effects in the respective count regression model.

3.2. Remarks on large sample results

The validity of asymptotic results depends on conditions that are specific to the assumed parametric distribution for count data, although the theory of maximum-likelihood estimation and mixed models provide guidance; see, e.g., Newey and McFadden (1994) and Demidenko (2004). In terms of ZIP models, Min and Czado (2010) introduce conditions and establish asymptotic results for maximum likelihood estimation, and He, Xue, and Shi (2010) derive results for sieve maximum likelihood estimation. The consistency of $\hat{\boldsymbol{\vartheta}}$ in (11) follows from large sample results for mixed effects models (Breslow and Clayton, 1993; Hui, Müller, and Welsh, 2017). Under Assumptions A1–A8 in Appendix C, the conditions of Proposition 2, and the continuous mapping theorem, $\hat{y}_{it}^\tau \xrightarrow{p} y_{it}^\tau$.

Finally, the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}^\tau$ in (15) follow from existing theoretical work, as we employ an NLMM model similar to the ones in Ibrahim, Zhu, Garcia, and Guo (2011) and Hui, Müller, and Welsh (2017). Furthermore, under regularity assumptions similar to the ones in Appendix C, including that the random effects are not allowed to grow, Theorem 2.5 and Theorem 3.3 in Newey and McFadden (1994) can be employed to obtain large sample results. We emphasize the main differences are relative to routine conditions, such as compactness of the parameter space, smoothness of the loglikelihood function, and conditions on the Fisher information matrix. Consistent with (13), $\epsilon_{it}^\tau := y_{it}^\tau - h(\eta_{it}^\tau)$ has zero conditional mean and bounded variance

for all i and t, and the inverse log link function $h(\cdot)$ and the distribution of y_{it} belong to a class of models satisfying the conditions of Propositions 1 and 2.

3.3. Statistical inference

The estimation of the asymptotic covariance matrix of multi-step methods can suffer from finite sample biases; see Chen, Chen, and Zhou (2004) and Windmeijer (2005) for similar problems. The estimated asymptotic variance of $\hat{\boldsymbol{\beta}}^{\tau}$ can exhibit similar biases, because it relies on the consistency of the first- and second-step estimators. This is an important practical consideration in a wide range of applications as it could lead to inaccurate inference and size distortions.

In this paper, we do not offer a finite-sample correction, but instead we turn to the bootstrap for statistical inference for the three-step estimator. The estimation is carried out by employing a bootstrap strategy similar to the one considered in Chernozhukov, Fernández-Val, and Weidner (2021). They propose the multiplier bootstrap in the case of a Poisson model. In our version of the bootstrap, we multiply the objective functions in the multiple steps by weights drawn from an exponential distribution with mean and variance equal to one. In results presented in the next section, we find that the coverage probabilities of the multiplier bootstrap confidence intervals tend to give results close to the nominal probabilities under the ZIP and the ZINB distributions. Moreover, using the bootstrap procedure, we provide confidence intervals for the empirical application in Section 5.

While this paper focuses on identification and estimation of 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).

4. Numerical Study

We next conduct a simulation study designed to evaluate the finite-sample performance of our method proposed in Section 3. We first present results for models with a fixed proportion of zero inflation, and then we include simulations for the case where the model generates a proportion of zeros that varies by subject and time. The estimates obtained for these models using the three-step estimator are benchmarked against the results obtained using jittering. Jittering is performed according to Section 3.6 of Machado and Santos Silva (2005), where a log-linear quantile regression model is fit using the jittered data. This is implemented using the *rq.counts()* function of the `R` package `Qtools` (Geraci, 2016).

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 is 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

N	T	β_j	ZIP		ZINB		ZIP		ZINB	
			Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
			$\pi_{it} = 0$				$\pi_{it} = 0.30$			
150	5	β_1	0.000	0.025	0.001	0.038	-0.002	0.034	0.000	0.051
		β_2	0.000	0.015	0.000	0.018	0.001	0.018	0.000	0.021
150	10	β_1	0.001	0.017	0.000	0.026	0.000	0.023	0.000	0.036
		β_2	0.000	0.014	0.000	0.015	0.000	0.015	0.000	0.015
250	5	β_1	0.000	0.020	0.001	0.030	0.000	0.026	0.000	0.040
		β_2	0.000	0.012	0.000	0.014	0.000	0.014	0.000	0.017
250	10	β_1	0.000	0.013	0.000	0.020	0.000	0.017	0.000	0.027
		β_2	0.000	0.011	0.000	0.012	0.000	0.012	0.000	0.013
			$\pi_{it} = 0.15$				Varying π_{it}			
150	5	β_1	-0.001	0.029	0.001	0.044	-0.001	0.030	0.001	0.039
		β_2	0.000	0.017	0.000	0.019	0.000	0.015	0.000	0.018
150	10	β_1	0.001	0.021	-0.001	0.030	0.000	0.020	0.000	0.026
		β_2	0.000	0.014	0.001	0.016	0.000	0.015	0.000	0.015
250	5	β_1	0.001	0.024	0.002	0.035	0.001	0.023	0.001	0.031
		β_2	0.000	0.014	0.000	0.015	0.000	0.013	0.001	0.015
250	10	β_1	0.000	0.016	-0.001	0.023	0.000	0.015	0.000	0.021
		β_2	0.000	0.011	0.000	0.012	0.000	0.011	0.000	0.012

Table 1: Bias and RMSE of β_1 and β_2 using the first-step estimator for the mean model.

$p_{y_{it}|\mathbf{u}_i}$ 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 $x_{it} = r_0 + r_1\zeta_i + r_2\zeta_{it}$, and the variables ζ_i and ζ_{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 $\tau \in \{0.50, 0.75, 0.90\}$. We evaluate the small sample performance of our approach by calculating the bias and root mean square error (RMSE) of the first-step and third-step estimators. The results for the parameters β_1 and β_2 of the conditional mean model estimated in the first step for the ZIP and ZINB models are given in Table 1. These biases and RMSEs of the mean effect are calculated with respect to the parameter values $\beta_1 = 0.25$ and $\beta_2 = 0$. Clearly, both estimates are quite accurate and practically unbiased, regardless of the amount of zero inflation and the count model used.

These first-step estimators for the conditional mean model are then used to produce the third-step estimators for β_1^τ and β_2^τ using the quantile response variable that was found as a solution of the nonlinear equation specified in the second step. However, the values corresponding to β_1^τ and β_2^τ are not known. The strategy we employ is to determine pseudo-true parameter values via simulation using large samples and considering the latent variable u_i as a regressor.

N	T	Method	$\tau = 0.50$		$\tau = 0.75$		$\tau = 0.90$	
			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
Varying π_{it}								
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

Table 2: Bias and RMSE of β_1^τ estimators when the response is distributed as ZIP. The jittering approach is denoted by JIT and the three-step approach is denoted by TS.

The values for $\{\beta_1^{\tau=0.50}, \beta_1^{\tau=0.75}, \beta_1^{\tau=0.90}\}$ are $\{0.265, 0.223, 0.197\}$ for the Poisson distribution and $\{0.268, 0.238, 0.225\}$ for the negative binomial distribution. As expected, β_2^τ is equal to zero at different quantiles and distributions. Albeit our strategy to calculate the pseudo-true parameter values 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 2 shows the small sample performance of the estimator for β_1^τ when y_{it} is distributed as ZIP. The top three-quarters of the table give the results when a constant proportion of zero inflation is assumed for the model. We consider $\pi_{it} \in \{0, 0.15, 0.30\}$, corresponding to no zero inflation, moderate zero inflation,

N	T	Method	$\tau = 0.50$		$\tau = 0.75$		$\tau = 0.90$	
			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
Varying π_{it}								
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: Bias and RMSE of β_1^τ estimators when the response is distributed as ZINB. The jittering approach is denoted by JIT and the three-step approach is denoted by TS.

and high zero inflation, respectively. The results for the three-step estimator are given in the rows labeled "TS" under the "Method" column. The results indicate that the method performs quite well, yielding negligible biases for the quantile effects. For each combination of subjects and time, the table shows negligible 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. We also compared our results to those obtained using jittering. These results are given in the corresponding rows labeled "JIT" under

the "Method" column. For each data generating process, the results for the three-step estimator are consistently better in terms of bias and RMSE than the jittering result. This provides clear empirical evidence of the improved estimation of the quantile effects obtained under our three-step estimator over the jittering approach. When we turn our attention to the case of ZINB shown in Table 3, we find similar results. The three-step estimator again performs well, the results do not seem to vary across the different proportions of zero inflation, and there is consistently better performance over the jittering estimator. Finally, similar conclusions in terms of the RMSE are also drawn about the estimator for β_2^τ, thus the results are not presented here to save space.

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:

$$\text{logit}(\pi_{it}) = \gamma_0 + \gamma_1 w_{it}, \tag{16}$$

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 the first step, these probabilities are estimated using maximum likelihood estimation, but are not reported to save space.

The bottom quarter of Tables 2 and 3 show the small sample performance of the estimators for the slope parameters when y_{it} is distributed as ZIP and ZINB, respectively, where π_{it} is generated according to Equation (16). Once again, excellent performance of the three-step estimator is found in each setting with noticeable improvement over the jittering estimator. These results also indicate 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. As expected, the bias of the estimator is small and the RMSE tends to decrease as the sample size increases.

Lastly, we offer results on the validity of the bootstrap procedure employed in the next section. Table 4 provides empirical coverage probabilities obtained by the bootstrap for a nominal 95% confidence interval. The probabilities are calculated based on asymptotic Gaussian confidence intervals. The coverage probabilities for β_1^τ are close to the nominal level of 0.95 in the case of the ZIP model, but they appear to be more liberal for the ZINB model. In most variants of the model, we observe a similar coverage performance across the different degrees of zero inflation.

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

N	T	π_{it}	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
			ZIP	ZINB	ZIP	ZINB	ZIP	ZINB
150	5	0.00	0.925	0.915	0.915	0.907	0.915	0.897
		0.15	0.940	0.910	0.938	0.890	0.935	0.887
		0.30	0.948	0.925	0.935	0.910	0.935	0.905
150	10	0.00	0.935	0.932	0.938	0.920	0.938	0.915
		0.15	0.938	0.935	0.930	0.945	0.922	0.943
		0.30	0.948	0.948	0.927	0.935	0.930	0.927
250	5	0.00	0.927	0.920	0.938	0.915	0.927	0.917
		0.15	0.943	0.922	0.940	0.910	0.940	0.905
		0.30	0.932	0.935	0.927	0.915	0.930	0.907
250	10	0.00	0.945	0.930	0.935	0.935	0.925	0.925
		0.15	0.932	0.935	0.932	0.930	0.930	0.930
		0.30	0.932	0.900	0.927	0.905	0.935	0.900

Table 4: Coverage probability for β_1^τ under different degrees of zero inflation when the response is distributed as ZIP and ZINB.

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 characterize 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.

5.1. 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. The vast majority of participants are observed either 3 or 5 times, and 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. These counts comprise the vector $\mathbf{y}_i$, $i = 1, \ldots, 5908$ for the GLMMs estimated below. 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. These covariates comprise the vector of independent variables $\mathbf{x}_{it}$. For detailed variable definitions and summary statistics, see Table 6 in Appendix D.

The use of a ZI count model is also supported by some features of the RHIE. As mentioned in Section 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 physician while others have multiple visits. Under this circumstance, a random zero count could be observed, but a positive integer-valued count is also possible.

5.2. Model specification

In the first step, we model the conditional mean considering four different specifications: Poisson, negative binomial, ZIP, and ZINB. We then investigate the best parametric model for the RHIE. We begin by assessing the randomized quantile residuals of the four models (Dunn and Smyth, 1996). The randomized quantile residual plot provides a visual illustration that can be used for assessing a (ZI) count regression model's fit. The comparison presented in Figure 2 reveals important differences between models. Clearly, the Poisson regression model yields the worst fit because it fails to capture the presence of zero-inflation and certain large values. ZIP regression shows some improvement over Poisson regression, but still fails to adequately capture the overdispersion. On the other hand, both negative binomial regression and ZINB regression provide satisfactory fits to the data. We then employed a boundary-corrected likelihood ratio test to evaluate ZINB versus negative binomial. We obtained a test statistic of 227, which is significant at standard levels and it favors the specification with zero inflation. Finally, to complement the analysis, we calculated the Akaike information criterion (AIC) and Bayesian information criterion (BIC) for each model fit, and the comparison of models favored the ZINB, consistent with our previous analysis and Deb and Trivedi (2002). The above results, combined with the existence of both zero inflation and a long tail to the right, suggest that the ZINB model provides the best fit.

In the second step, we estimate a ZINB regression 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 $\mathbf{w}_{it}$ that includes 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. Thus, the design variable to predict the random effects will simply be $z_{it} = 1$, and z_{it} and the vector

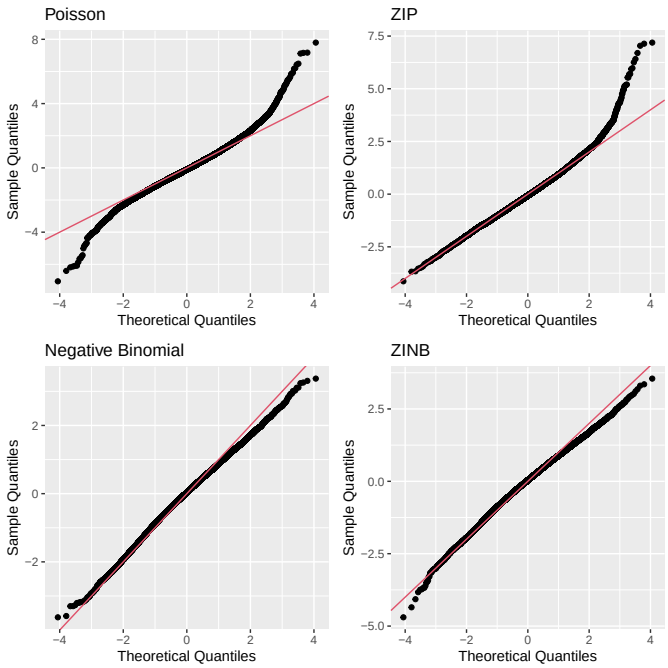

Figure 2: Randomized quantile residuals for the estimated Poisson, zero-inflated Poisson (ZIP), negative binomial, and zero-inflated negative binomial (ZINB) regression models.

of predictor variables $\mathbf{x}_{it}$ will be used to define the linear predictor η_{it} as defined in (6). Modeling individual specific intercepts as random effects is consistent with the use of experimental data, and a key 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.

5.3. Empirical Results

Table 5 presents results for 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 0.50, 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 5 are similar to the estimates in Table 4 in Deb and Trivedi (2002), although the estimates presented here are estimated more precisely. The sign of the coefficients are 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

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)	0.0000 (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)

Table 5: Estimated regression coefficients, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{\tau}$, for the RAND Health Insurance Experiment dataset. Standard errors are in parentheses.

count variable to be negative because the cost shared by the patient 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 observe 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 are quantitatively similar to the estimated effects at the 0.75 and 0.90 quantiles, revealing not only the importance of distributional effects, but also that the mean effect offers an incomplete description of the effect of some insurance, demographic, and socioeconomic variables. To examine this claim in more detail, we estimate the model as in Table 5, but now considering 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 5.

Figure 3 shows the estimated mean effects (dashed lines) and quantile effects (continuous lines) obtained from our proposed method. In order to examine the importance of accommodating 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 (2019). The figure reveals some interesting new findings. First, we find that the health insurance option associated with coinsurance (LC) significantly reduce medical care utilization, particularly among 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

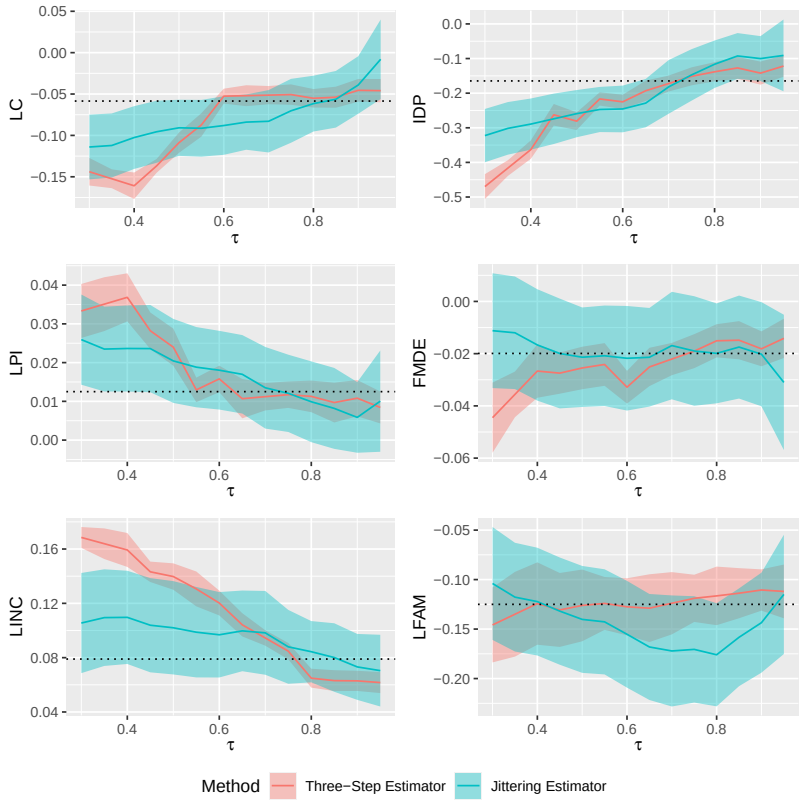

Figure 3: *Estimated regression coefficients, $\hat{\boldsymbol{\beta}}^{\tau}$, for policy variables, socio-economic variables, and demographic variables. The dotted line is the estimated value of $\boldsymbol{\beta}$ for the mean structure, obtained by a ZI GLMM model in the first step with ZINB specification.*

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.

6. Discussion

The primary aim of this work is to study identification and estimation of conditional quantile functions for count responses with zero inflation in the longitudinal data setting. Our approach is based on using a continuous approximation to the discrete distribution for the count model under consideration. This approach has been developed in Ilienko (2013) and Padellini and Rue (2019), with

the latter leveraging this approximation to perform quantile regression for count data. We extended to the longitudinal setting where the count responses are also subject to zero inflation. Another important distinction from Padellini and Rue (2019) is that we first consider estimation of the conditional mean rather than considering a quantile regression model. This important 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 are ZI GLMMs, which affords the practitioner considerable flexibility regarding the structural form of the model for their application. The class of ZI GLMMs are, 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 some, or all, of 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 unknown dispersion parameter. Having this broad class of distributions at our disposal allows for effective exploration of reasonable and meaningful structures to consider for the conditional mean structure for the application at hand. Maximum likelihood estimation is accomplished using the Laplace approximation to calculate the marginal likelihood, which is able to 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 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 4 that the proposed estimator has satisfactory performance for the estimation of 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 through examination of quantile effects while capturing subject heterogeneity as well as 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 about how certain 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 novel approach to understand quantile effects when modeling ZI longitudinal count responses.

Acknowledgments

We are grateful to the Editor, an Associate Editor, and two anonymous reviewers for helpful comments and suggestions, which significantly improved our paper. We would also like to thank Matt Harding, Solomon Harrar, Arnold Stromberg, and Chenglong Ye for comments and useful conversations.

Appendix A. Proofs

Proof of Proposition 1. 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, 2013; Padellini and Rue, 2019).

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) and (4) satisfy the following:

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

and

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

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. □

Proof of Proposition 2. The existence of the solution can be shown as a direct application of Bolzano's theorem. Let $K_{it}(y, \tau) = a_{it} - b_{it}k(y, \theta_{it})$, where $a_{it} = \tau - \pi_{it} > 0$, and $b_{it} = 1 - \pi_{it} > 0$. Because $k(y, \theta_{it})$ is an increasing function over $[0, 1]$, $K_{it}(0, \tau) = a_{it} > 0$, and $\lim_{y \to \infty} K_{it}(y, \tau) = \tau - 1 < 0$. Therefore, we have two points y_0 and y_1 such that $K_{it}(y_0, \tau) > 0$ and $K_{it}(y_1, \tau) < 0$, and thus a root exists. The uniqueness result follows because $k(y, \theta_{it})$ is increasing in y. □

Appendix B. Implementation details

In the PNLS step, the current estimate of $\boldsymbol{\Delta}^\tau$ is held fixed. The conditional modes of the random effects $\mathbf{u}_i^\tau$ and the conditional estimates of the fixed effects $\boldsymbol{\beta}^\tau$ are obtained by minimizing a PNLS objective function:

$$\hat{\boldsymbol{\beta}}^\tau = \underset{\{\boldsymbol{\beta}^\tau | \boldsymbol{\Delta}^\tau\}}{\arg\min} \sum_{i=1}^{N} \left\{ \|\mathbf{y}_i^\tau - h(\boldsymbol{\eta}_i^\tau)\|^2 + \|\boldsymbol{\Delta}^\tau \mathbf{u}_i^\tau\|^2 \right\}. \tag{17}$$

The LMM step updates the estimate of $\boldsymbol{\Delta}^\tau$ based on a first-order Taylor expansion of the model function $h(\cdot)$ around the current estimates of $\boldsymbol{\beta}^\tau$ and the conditional modes of the random effects $\mathbf{u}_i^\tau$, which we will denote by $\hat{\boldsymbol{\beta}}^{\tau(w)}$ and $\hat{\mathbf{u}}_i^{\tau(w)}$, respectively. Let

$$\hat{\mathbf{X}}_i^{(w)} = \frac{\partial h(\boldsymbol{\eta}_i^\tau)}{\partial \boldsymbol{\beta}^{\tau\top}}\bigg|_{\hat{\boldsymbol{\beta}}^{\tau(w)}, \hat{\mathbf{u}}_i^{\tau(w)}}, \quad \hat{\mathbf{Z}}_i^{(w)} = \frac{\partial h(\boldsymbol{\eta}_i^\tau)}{\partial \mathbf{u}^{\tau\top}}\bigg|_{\hat{\boldsymbol{\beta}}^{\tau(w)}, \hat{\mathbf{u}}_i^{\tau(w)}},$$
$$\hat{\mathbf{v}}_i^{\tau(w)} = \mathbf{y}_i^\tau - h(\boldsymbol{\eta}_i^{\tau(w)}) + \boldsymbol{\eta}_i^{\tau(w)},$$

where

$$\boldsymbol{\eta}_i^{\tau(w)} = \hat{\mathbf{X}}_i^{(w)}\hat{\boldsymbol{\beta}}^{\tau(w)} + \hat{\mathbf{Z}}_i^{(w)}\hat{\mathbf{u}}_i^{\tau(w)}.$$

The quantity $\hat{\mathbf{v}}_i^{\tau(w)}$ is a modified response vector with fixed-effects and random-effects design matrices $\hat{\mathbf{X}}_i^{(w)}$ and $\hat{\mathbf{Z}}_i^{(w)}$, respectively. Finally, let $\boldsymbol{\Sigma}_i(\boldsymbol{\Delta}^\tau) = \mathbf{I}_T + \hat{\mathbf{Z}}_i^{(w)}(\boldsymbol{\Delta}^\tau)^{-1}(\boldsymbol{\Delta}^\tau)^{-\top}\hat{\mathbf{Z}}_i^{(w)\top}$. The approximate loglikelihood then used to estimate $\boldsymbol{\Delta}^\tau$ is

$$\ell^*(\boldsymbol{\beta}^\tau, \sigma^{2\tau}, \boldsymbol{\Delta}^\tau|\mathbf{y}^\tau) = -\frac{N}{2}\log(2\pi\sigma^{2\tau}) - \frac{1}{2}\sum_{i=1}^{N}\Bigg\{\log|\Sigma_i(\boldsymbol{\Delta}^\tau)| + \sigma^{-2\tau}\left[\hat{\mathbf{v}}^{(w)} - \hat{\mathbf{X}}_i^{(w)}\boldsymbol{\beta}^\tau\right]^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}^\tau)\left[\hat{\mathbf{v}}^{(w)} - \hat{\mathbf{X}}_i^{(w)}\boldsymbol{\beta}^\tau\right]\Bigg\}.$$

The above PNLS routine just described is implemented using the `nlmer()` function. Note that a restricted maximum likelihood estimator for $\boldsymbol{\Delta}^\tau$ can also be found by using the following restricted loglikelihood:

$$\ell_R^*(\boldsymbol{\beta}^\tau, \sigma^{2\tau}, \boldsymbol{\Delta}^\tau|\mathbf{y}^\tau) = \ell^*(\boldsymbol{\beta}^\tau, \sigma^{2\tau}, \boldsymbol{\Delta}^\tau|\mathbf{y}^\tau) - \frac{1}{2}\sum_{i=1}^{N}\log\left|\sigma^{-2\tau}\hat{\mathbf{X}}_i^{(w)\top}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}^\tau)\hat{\mathbf{X}}_i^{(w)}\right|.$$

Appendix C. Regularity Conditions

In this appendix, we outline the necessary regularity conditions for the consistency and asymptotic normality of $\hat{\boldsymbol{\vartheta}}$ and $\hat{\boldsymbol{\beta}}^\tau$. For the asymptotic results on $\hat{\boldsymbol{\vartheta}}$, we adapt the regularity conditions found across Newey and McFadden (1994), Gan (2000), and Demidenko (2004) for our ZI GLMM setting. For ease of notation in what follows, we will let $f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta}) \equiv f_{y_{it}|\mathbf{u}_i}(y_{it};\mathbf{x}_{it},\mathbf{w}_{it},\mathbf{z}_{it},\mathbf{u}_i,\boldsymbol{\vartheta})$ for the ZI GLMM pmf given in Equation (8). Moreover, let $d^* = \dim(\boldsymbol{\vartheta})$, $\mathcal{C}^l$ be the space of continuous functions with l continuous derivatives, and $\mathcal{B}_a(\boldsymbol{\delta})$ be the open ball of radius $a > 0$ centered at the point $\boldsymbol{\delta}$.

- A1. For $y \in \mathbb{N}$ and each $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}$, $f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta}) > 0$, where $\boldsymbol{\Theta}$ is a connected subset in $\mathbb{R}^{d^*}$.
- A2. For each $y \in \mathbb{N}^+$, $f_{y|\mathbf{u}}(y;\cdot)$ as a function of $\boldsymbol{\vartheta}$ is the constant 0 on the boundary $\partial\boldsymbol{\vartheta}$ of $\boldsymbol{\Theta}$, and it is in $\mathcal{C}^2$ on $\boldsymbol{\Theta}$; i.e., the second derivative of $f_{y|\mathbf{u}}(y;\cdot)$ is continuous on $\boldsymbol{\Theta}$.
- A3. For each $y \in \mathbb{N}$, $f_{y|\mathbf{u}}(y;\cdot)$ as a function of $\boldsymbol{\vartheta}$ is in $\mathcal{C}^0$, $\mathcal{C}^1$, and $\mathcal{C}^2$ on $\boldsymbol{\Theta}$; i.e., $f_{y|\mathbf{u}}(y;\cdot)$ and its first two derivatives are each continuous on $\boldsymbol{\Theta}$.
- A4. For each $y \in \mathbb{N}^+$, $f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta})$ as a function of $\boldsymbol{\vartheta}$ is in $\mathcal{C}^3$ on $\boldsymbol{\Theta}$.
- A5. If for each pair of $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ there is a constant $c(\boldsymbol{\vartheta}_1,\boldsymbol{\vartheta}_2)$, which might depend on $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$, such that $f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta}_1) = c(\boldsymbol{\vartheta}_1,\boldsymbol{\vartheta}_2)f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta}_2)$ for $y \in \mathbb{N}^+$, then $\boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_2$.
- A6. For $\boldsymbol{\vartheta}_0$, the true parameter, there exists $a > 0$ such that $\mathcal{B}_a(\boldsymbol{\vartheta}_0) \subset \boldsymbol{\Theta}$ and the following hold:

(a) For each $y \in \mathbb{N}^+$, $f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta})$ as a function of $\boldsymbol{\vartheta}$ is in $\mathcal{C}^1$ and $\mathcal{C}^2$ on $\mathcal{B}_a(\boldsymbol{\vartheta}_0)$.

(b) $\sum_{i=0}^{\infty} \left\| \frac{\partial f_{y|\mathbf{u}}(i;\boldsymbol{\vartheta}_0)}{\partial \boldsymbol{\vartheta}} \right\| < \infty$.

(c) There is a sequence of nonnegative numbers, M_i, $i = 0, 1, \ldots$, such that $\sum_{i=0}^{\infty} M_i < \infty$ and for any $\boldsymbol{\vartheta} \in \mathcal{B}_a(\boldsymbol{\vartheta}_0)$, $\left\| \frac{\partial f_{y|\mathbf{u}}(0;\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \right\| \leq M_i$ and $\left\| \frac{\partial f_{y|\mathbf{u}}(0;\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^\top} \right\| \leq M_i$.

A7. Let y have the ZI GLMM pmf $f_{y|\mathbf{u}}(\cdot;\boldsymbol{\vartheta}_0)$. Then, the following are true:

(a) $\mathrm{E}_{\boldsymbol{\vartheta}_0} \left\| \frac{\partial \log f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta}_0)}{\partial \boldsymbol{\vartheta}} \right\|^2 < \infty$.

(b) The information matrix $I_{\boldsymbol{\vartheta}_0} = \mathrm{Var}_{\boldsymbol{\vartheta}_0} \left[\frac{\partial \log f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta}_0)}{\partial \boldsymbol{\vartheta}} \right]$ is positive definite.

A8. There is a positive number a and a nonnegative measurable function on the set of nonnegative integers, $M(\cdot)$, where $\mathrm{E}_{\boldsymbol{\vartheta}_0} M(y) < \infty$ such that for any $\boldsymbol{\vartheta} \in \mathcal{B}_a(\boldsymbol{\vartheta}_0)$,

$$\left| \frac{\partial^3 \log f_{y|\mathbf{u}}(y;\boldsymbol{\vartheta})}{\partial \vartheta_i \partial \vartheta_j \partial \vartheta_k} \right| \leq M(y), \quad \text{for } i, j, k = 1, \ldots, d^*.$$

Note that the previous assumptions imply Assumption (C1) in Ibrahim, Zhu, Garcia, and Guo (2011). Under the previous conditions and regularity conditions on the penalty function and tuning parameter, Theorem 1 in Ibrahim, Zhu, Garcia, and Guo (2011) states the consistency and asymptotic normality of the maximum penalized likelihood estimator for mixed-effects models.

Appendix D. Definitions and summary statistics for the RHIE data

Variables	Definition	Mean	Min	Quantile			Max
				0.25	0.50	0.75	
MDU	Yearly number of outpatient visits to physicians	2.86	0	0	1	4	77
LC	ln(coinsurance+1), $0 \leq$ coinsurance rate ≤ 100	2.38	0	0	3.26	4.56	4.56
IDP	Indicator for individual deductible plan	0.26	0	0	0	1	1
LPI	ln(max(1,annual participation incentive payment))	4.71	0	4.06	6.11	6.62	7.16
FMDE	log(max(medical deductible expenditure))	4.03	0	0	6.09	6.96	8.29
PHYSLIM	Indicator for physical limitations	0.12	0	0	0	0	1
NDISEASE	Index of chronic diseases	11.24	0	6.90	10.58	13.73	58.60
LINC	ln(annual family income) in US dollars	8.71	0	8.58	8.98	9.26	10.28
LFAM	ln(family size)	1.25	0	1.10	1.39	1.61	2.64
EDUCDEC	Education of head of household in years	11.97	0	11	12	13	25
AGE	Age in years	25.72	0	11.46	24.19	37.40	64.28
FEMALE	Indicator for female	0.52	0	0	1	1	1
CHILD	Indicator for age less than 18	0.40	0	0	0	1	1
FEMCHILD	FEMALE*CHILD	0.19	0	0	0	0	1
BLACK	1 if race of household head is black	0.18	0	0	0	0	1
HLTHG	1 if self-rated health is good	0.36	0	0	0	1	1
HLTHF	1 if self-rated health is fair	0.08	0	0	0	0	1
HLTHP	1 if self-rated health is poor	0.02	0	0	0	0	1

Table 6: Variable definitions and summary statistics for the RAND Health Insurance Experiment dataset.

References

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

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

M. L. Battagliola, H. Sørensen, A. Tolver, and A. M. Staicu. A Bias-Adjusted Estimator in Quantile Regression for Clustered Data. *Econometrics and Statistics (in press)*, 2021. doi: https://doi.org/10.1016/j.ecosta.2021.07.003.

F. J. Breidt. Simulation Estimation of Quantiles from a Distribution with Known Mean. *Journal of Computational and Graphical Statistics*, 13(2):487–498, 2004.

N. E. Breslow and D. G. Clayton. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

M. E. Brooks, K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Mächler, and B. M. Bolker. glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling. *The R Journal*, 9(2):378–400, 2017. URL https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

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

J. Chen, J. Chen, and H. Zhou. Two-Step Estimation for a Generalized Linear Mixed Model with Auxiliary Covariates. *Statistica Sinica*, 14(2):361–376, 2004.

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

V. Chernozhukov, I. Fernández-Val, and M. Weidner. Network and Panel Quantile Effects via Distribution Regression. Journal of Econometrics, forthcoming, 2021.

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

E. Demidenko. *Mixed Models: Theory and Applications*. Wiley Series in Probability and Statistics, 2004.

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

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

A. R. Gallant and D. W. Nychka. Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica*, 55(2):363–390, 1987.

N. Gan. *Generalized Zero-Inflated Models and Their Applications*. PhD thesis, North Carolina State University, 2000.

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

M. Geraci and M. Bottai. Linear Quantile Mixed Models. *Statistics and Computing*, 24(3):461–479, 2014.

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

X. He, H. Xue, and N.-Z. Shi. Sieve Maximum Likelihood Estimation for Doubly Semiparametric Zero-Inflated Poisson Models. *Journal of Multivariate Analysis*, 101(9):2026–2038, 2010.

J. M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK, 2$^{\text{nd}}$ edition, 2011.

F. K. C. Hui, S. Müller, and A. H. Welsh. Hierarchical Selection of Fixed and Random Effects in Generalized Linear Mixed Models. *Statistica Sinica*, 27 (2):501–518, 2017.

J. G. Ibrahim, H. Zhu, R. I. Garcia, and R. Guo. Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*, 67(2):495–503, 2011.

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

T. Kneib, A. Silbersdorff, and B. Säfken. Rage Against the Mean A Review of Distributional Regression Approaches. *Econometrics and Statistics (in press)*, 2021. doi: https://doi.org/10.1016/j.ecosta.2021.07.006.

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

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

K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell. TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70(5):1–21, 2016. URL https://www.jstatsoft.org/article/view/v070i05.

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

D. Lee and T. Neocleous. 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, 2010.

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

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

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

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

A. Min and C. Czado. Testing for Zero-Modification in Count Regression Models. *Statistica Sinica*, 20(1):323–341, 2010.

W. K. Newey and D. McFadden. Large Sample Estimation and Hypothesis Testing. volume 4 of *Handbook of Econometrics*, chapter 36, pages 2111 – 2245. 1994.

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

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

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

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

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

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

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

F. Windmeijer. A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators. *Journal of Econometrics*, 126(1):25–51, 2005.

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

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

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

D. Zhang and M. Davidian. Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data. *Biometrics*, 57(3):795–802, 2001.

H. Zhu, S. Lou, and S. M. DeSantis. Zero-Inflated Count Models for Longitudinal Measurements with Heterogeneous Random Effects. *Statistical Methods in Medical Research*, 26(4):1774–1786, 2017.