# Zero-inflated Poisson regression model for a new class of flexible link functions: A case study on healthcare utilization

Essoham Ali

▶ **To cite this version:**

Essoham Ali. Zero-inflated Poisson regression model for a new class of flexible link functions: A case study on healthcare utilization. 2020. hal-03091927

**HAL Id: hal-03091927**
**https://hal.science/hal-03091927v1**

Preprint submitted on 31 Dec 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Zero-inflated Poisson regression model for a new class of flexible link functions: A case study on healthcare utilization

Essoham ALI

*University Gaston Berger, LERSTAD, Saint-Louis, Senegal*

**Abstract**

Many disciplines produce count data that contain many zeros. Zero inflation count models such as ZIP and ZIB have been widely used to model count data, in particular, to model the latent structure in a ZIP regression model that allows a non-linear functional relationship between covariates and the expected count outcome. A critical problem in modeling the count response data is the appropriate choice of links functions. Commonly used link functions such as logit link have fixed skewness but lack in flexibility to allow the data to determine the degree of the skewness. To overcome this limitation, we propose a flexible ZIP regression model that combines a generalized extreme value link function with the other link functions. The maximum likelihood estimator is used in the estimation of the model. Maximum likelihood estimation is effective in this model in a series of scenarios. Through simulated data sets and analysis of the healthcare utilization, we show that the proposed link function is quite flexible and performs better against link misspecification than standard link functions.

*Keywords:* Flexibility; Excess of zeros; Generalized extreme value distribution; Health-care demand; Simulations

## 1. Introduction

Statistical modelling is an important step of data analysis in many fields of scientific research or decision-making purpose. To carry out this approach, one needs to specify a probability distribution that accounts as accurate as possible the variability observed in data. Given the plethora of discrete or continuous distributions available (e.g., Johnson et al. , 2005), guidelines are needed to identify not randomly the one or two-parameter family of distributions suited for modelling data on hand. The practice in this procedure is to consider in addition certain phenomenon such as: over-/under-dispersion or zero inflation/deflation for count data ( (e.g., Bonat et al. , 2018)) and, over-/under-variation or zero mass for continuous data ( (e.g., Abid et al., 2020)).

*Email address:* ali.essoham@ugb.edu.sn (Essoham ALI)
corresponding author: Essoham ALI

The Poisson dispersion phenomenon is well-known and very widely used in practice; see, (e.g., Kokonendji, C.C. ) for a review of count (or discrete integer-valued) models. Various models have been developed to address zero-inflation, such as zero-inflated (ZI) models which mix a degenerate distribution at zero with a standard count model. Zero-inflated Poisson (ZIP) regression model was proposed by Lambert (1992) and further developed by Dietz et al. (2000), Lim et al. (2006) and Monod (2014), among many others. Recent variants of ZIP regression include random-effects ZIP models (Hall (2000); Min and Agresti (2005)) and semiparametric ZIP models Lam et al. (2006). A Zero-inflated negative binomial (ZINB) regression model was proposed by Ridout et al. (2001), see also Moghimbeigi et al. (2008).

Thus,Hall (2000) introduced the zero-inflated binomial (ZIB) model, see also Hall and Berenhaut (2002), Diop et al. (2011), and Diallo et al. (2017). Deng and Zhang (2015) proposed a zero-one inflated binomial regression model for such data. In Nguyen et al. (2019), authors proposed a Zero-inflated Poisson regression with right-censored data. The usual way to model the response variable is to use a Generalized Linear Model (GLM), where we model the latent probability of "success" by a linear function of covariates through a link function McCullagh et al. (1989). The logit, probit and Student t link functions are three of the common links used in GLM. However, the link functions mentioned above are "symmetric" links in the sense that they assume that the latent probability of a given response variable approaches 0 with the same rate as it approaches 1. Equivalently, the probability density function that corresponds to the inverse cumulative distribution function of the link function is symmetric. However, this may not be a reasonable assumption in many cases. One commonly adopted asymmetric link function is the complementary loglog (cloglog) link function. However, the cloglog link has a fixed negative skewness. As a result, it lacks both the flexibility to let the data display how much skewness should be incorporated and also the ability to allow positive skewness. In short, count data might often be better modeled with flexible link functions that allow for both positive and negative skewness and that allow the data to determine the amount of skewness required.

Many research works have been conducted which introduce flexibility into the link functions. Aranda-Ordaz (1981) proposed two separate one-parameter models for additional flexibility in the logistic model.Guerrero et al. (1982) used Box-Cox transformation on the odds ratio to form a more flexible class of model. Jones (2004) proposed a family of flexible distributions based on the distribution of order statistics. Stukel (1988) proposed a two-parameter class of generalized logistic models. Stukel's model approximates many standard symmetric and asymmetric link functions quite well, but in a Bayesian framework, it may result in improper posteriors when the usual improper uniform prior is used in regressions Chen et al. (1999) . Recently, Wang et al. (2010) proposed the generalized extreme value link function giving more flexible skewness controlled by the shape parameter. But the standard logistic and probit links are not among the special cases of this family.

A critical problem in modeling the count response data is the appropriate choice of links functions. To overcome this limitation, we propose a flexible Zero-Inflated Poisson regression model that combines a generalized extreme value link function with the other link functions. In the extreme value theory, the GEV distribution is used to model the

tail of a distribution Coles S. G. (2004). Currently, the logistic regression model, with its convenient interpretation and implementation, has been routinely employed to estimate and predict.As in this work, we focus on the Poisson parameters we have chosen to vary several link functions in order to see the flexibility of the GEV distribution with respect to the others. In the GLM, Agresti, A. (2002), log-log and complementary log-log link functions are used since they are asymmetric functions. In particular, the log-log link function is the quantile function of the Gumbel random variable. The inverse function of the complementary log-log is equal to one minus the cumulative distribution function of the Gumbel random variable. Consequences of link misspecification have been studied by numerous authors in the literature. In particular, for independent binary observations, Czado et al. (1992) demonstrated that falsely assuming a logistic link leads to a substantial increase in the bias and the mean squared error of the parameter estimates as well as the predicted probabilities, both asymptotically and in finite samples. Moreover, these undesirable effects have greater magnitude when the misspecification involves skewness than when it involves kurtosis (or tail weight). Wu et al. (2002) showed also that under certain conditions there exist linear relationships between the regression coefficients though the choice of links is important for goodness of fit. To build an appropriate and extremely flexible model for the count data and to overcome the constraint for the skewed generalized link models, we propose the cloglog, probit and generalized extreme value (GEV) distribution as a link function. In this paper we then suggest a new class of link functions to model count data, and apply it to healthcare utilization data. This paper is organized as follows. In Section 2, we recall the defnition of ZIP regression model, we describe the maximum likelihood estimation under different link functions and we introduce some useful notations. In Section 3, we report the results of our simulations. An application to a health-care utilization dataset is described in Section 4. Some concluding remarks are given in Section 5.

## 2. Notations and likelihood calculation

Let us first specify the notation used throughout this paper. Suppose $Z_i \sim \text{ZIP}(\lambda_i, \pi_i)$, where $\pi_i$ is the probability of success for the ith observation and $\lambda_i = e^{\beta^\top \mathbf{X}_i}$. Let the design matrix be $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{W}_i = (1, W_{i2}, \cdots, W_{iq})^\top$ are random vectors of predictors or covariates (both categorical and continuous covariates are allowed). The covariates vectors (they may share common components or be distinct), with $X_{i1} = W_{i1} = 1$. $\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^q$ are unknown parameter and $\top$ denotes the transpose operator. Let $J_i = 1_{\{Z_i=0\}}$. We associate $\pi_i$ and $\mathbf{W}_i$ through a cumulative distribution function $F$ as follows:

$$\pi_i = F(\gamma^\top \mathbf{W}_i) \tag{2.1}$$

where $F$ is a cumulative distribution function and $F^{-1}$ determines the link function.

*2.1. Susceptibility probability function with different links functions*

*2.1.1. Zero-inflated ZIP regression model*

The ZIP model assumes that the response variable $Z_i$ (where the lower indice i indicates the individual) is such that

$$Z_i \sim \begin{cases} 0 & \text{with probability } \pi_i, \\ \mathcal{P}(\lambda_i) & \text{with probability } 1-\pi_i, \end{cases} \tag{2.2}$$

where $\mathcal{P}(\lambda_i)$ denotes Poisson distribution with parameter $\lambda_i > 0$. Obviously, the ZIP model reduces to a standard Poisson distribution if $\pi_i = 0$. In ZIP regression, the mixing probability $\pi_i$ and parameter $\lambda_i$ are usually modeled by logistic and log-linear models respectively, that is:

$$F^{-1}(\pi_i) = \text{logit}(\pi_i) = \gamma^\top \mathbf{W}_i \tag{2.3}$$

and

$$\log(\lambda_i) = \beta^\top \mathbf{X}_i, \tag{2.4}$$

Suppose that we observe a sample of $n$ independent copies $(Z_i, \mathbf{X}_i, \mathbf{W}_i), i = 1, \ldots, n$ of $(Z, \mathbf{X}, \mathbf{W})$. For $i = 1, \ldots, n$, the log-likelihood of $\theta = (\beta^\top, \gamma^\top)^\top$ in the latent class ZIP model (2.6)-(2.3)-(2.4) is :

$$\begin{aligned} \ell_n(\theta) \;=\; \sum_{i=1}^{n} \Big\{ & J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + e^{-\exp(\beta^\top \mathbf{X}_i)}\right) + (1-J_i)\left[Z_i \beta^\top \mathbf{X}_i - \log\left(Z_i!\right)\right] \\ & - \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) \Big\}. \end{aligned}$$

The maximum likelihood estimator of $(\beta, \gamma)$ is obtained by maximizing this function. The ML estimator is consistent and asymptotically normally distributed (see Czado and Min (2005)).

*2.1.2. ZIP-GEV regression model*

A key component of the model given in (2.3)-(2.4) is the specification of the link function. The commonly used logit link is specified as $\pi_i = F(\gamma^\top \mathbf{W}_i)$, where $F$ is a cumulative distribution function (cdf) and $F^{-1}$ determines the link function. The symmetry in the normal distribution leads to the symmetry in the logit link. Wang et al. (2010) showed that the symmetric link has an inferior performance when the data structure requires a skewed response probability function. They proposed a link function based on the GEV distribution. The distribution function of GEV $(\mu, \sigma, \xi)$ is given by:

$$G(x|\mu,\sigma,\xi) = \begin{cases} \exp\left[-\left\{1+\xi\frac{(x-\mu)}{\sigma}\right\}_+^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left\{-\exp(\frac{(x-\mu)}{\sigma})\right\}, & \xi = 0, \end{cases} \tag{2.5}$$

where $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$ and $\xi \in \mathbb{R}$ are, respectively, the location,scale and shape parameters, and $x+ = \max(0, x)$. The shape of this distribution function is very flexible with the tail behavior controlled by the shape parameter $\xi$. When $\xi = 0$, it is the Gumbel distribution and decays exponentially. When $\xi < 0$, it reduces to the negative Weibull distribution with a finite short upper endpoint. When $\xi > 0$, it becomes the Fréchet distribution with a heavy tail behavior. The GEV link is the inverse of F which is assumed as

$$\pi_i = F(\mathbf{W}_i|\xi) = 1 - \text{GEV}(-\gamma^\top \mathbf{W}_i; \xi) = \begin{cases} 1 - \exp\left\{-(1 - \xi\gamma^\top \mathbb{W}_i)_+^{-1/\xi}\right\}, & \xi \neq 0, \\ 1 - \exp\left\{-\exp(\frac{(x-\mu)}{\sigma})\right\}, & \xi = 0, \end{cases} \tag{2.6}$$

where $\text{GEV}(x; \xi)$ represents the cumulative probability at $x$ for the GEV distribution with parameters $\phi = (\mu = 0, \sigma = 1, \xi)$. Note $\mu$ and $\sigma$ are set to fixed constants for model identifiability. Wang et al. (2010) showed that the GEV link model specified in (6) is negatively skewed for $\xi < \log 2 - 1$ and positively skewed for $\xi > \log 2 - 1$. The link function is approximately symmetric at $\xi = \log 2 - 1$. The cloglog link, specified as $F^{-1}(\pi_i) = -\log(-\log(\pi_i)) = \gamma^\top \mathbf{W}_i$, is a special case of the GEV link with $\xi = 0$.

The GEV regression model proposed by Calabrese et al. (2013) is defined by a link function that corresponds to the inverse cumulative function of the GEV distribution, that can be called GEV regression model or "gevit, in analogy with the "logit". The ZIP regression model under the GEV link is then given by

$$\text{gevit}(\pi_i) = \frac{[-\log(\pi_i)]^{-\xi} - 1}{\xi} = \gamma^\top \mathbf{W}_i = \gamma_1 + \sum_{j=1}^{q} \gamma_j \mathbf{W}_{ij} \tag{2.7}$$

$$\log(\lambda_i) = \beta^\top \mathbf{X}_i = \beta_1 + \sum_{k=1}^{p} \beta_k \mathbf{X}_{ik} \tag{2.8}$$

, $\beta \in \mathbb{R}^P$ , $\gamma \in \mathbb{R}^q$ are unknown parameters and $\top$ denotes the transpose operator and $\xi \in \mathbb{R}$ is the shape parameter for GEV distribution. According to (2.6)-(2.7)-(2.8), the log-likelihood of $\theta = (\beta^\top, \gamma^\top)^\top$

$$\begin{aligned} \ell_n^{GEV}(\theta) \;=\; & \sum_{i=1}^{n} \Big\{ J_i \log \left[ \exp\left[ -(1 + \xi\gamma^\top \mathbf{W}_i)^{-\frac{1}{\xi}} \right] + (1 - \exp\left[ -(1 + \xi\gamma^\top \mathbf{W}_i)^{-\frac{1}{\xi}} \right] ) e^{-\exp(\beta^\top \mathbf{X}_i)} \right] \\ & + (1 - J_i) \left[ Z_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} + \log\left( 1 - \exp\left[ -(1 + \xi\gamma^\top \mathbf{W}_i)^{-\frac{1}{\xi}} \right] \right) - \log(Z_i!) \right] \Big\}. \end{aligned}$$

The MLE $\hat{\theta}_n = (\hat{\beta}_n^\top, \hat{\gamma}_n^\top)^\top$ of $\theta$ is obtained by solving the score equation

$$\frac{\partial \ell_n^{GEV}(\theta)}{\partial \theta} = 0, \tag{2.9}$$

which can be achieved by nonlinear optimization

*2.1.3. ZIP-cloglog regression model*

The asymmetric cloglog link is specified as

$$F^{-1}(\pi_i) = -\log(-\log(\pi_i)) = \gamma^\top \mathbf{W}_i. \tag{2.10}$$

Assume that we observe n independent vectors $(Z_1, \mathbf{X}_1, \mathbf{W}_1), \cdots, (Z_n, X_n, W_n)$ from the model (2.4)-(2.10), all defined on the probability space $(\Omega, \mathcal{C}, \mathbb{P})$. The log-likelihood of $\theta = (\beta^\top, \gamma^\top)^\top$ based on these observations is

$$\begin{aligned}
\ell_n^{cloglog}(\theta) &= \sum_{i=1}^{n} J_i \log\left[e^{-\exp(-\gamma^\top \mathbf{W}_i)} + (1 - e^{-\exp(-\gamma^\top \mathbf{W}_i)})e^{-\exp(\beta^\top \mathbf{X}_i)}\right] \\
&\quad + \sum_{i=1}^{n}(1 - J_i)\left[Z_i\beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} + \log\left(1 - e^{-\exp(-\gamma^\top \mathbf{W}_i)}\right) - \log(Z_i!)\right], \\
&= \sum_{i=1}^{n} \ell_i(\theta).
\end{aligned}$$

The maximum likelihood estimator $\hat{\theta}_n = (\hat{\beta}_n^\top, \hat{\gamma}_n^\top)^\top$ of $\theta$ is solution of the $k-$ dimensional score equation

$$\dot{\ell}_i(\theta) = \frac{\partial \ell_n^{cloglog}(\theta)}{\partial \theta} = 0 \tag{2.11}$$

where $k = p + q$.

*2.1.4. ZIP-probit regression*

The zero-inflated Poisson model using the probit link function can be defined in the same way as the ZIP model, where the probability of zero inflation is modelled by the probit model. When risk factors are available, the mixing probability $\pi_i$ is usually modeled by a probit model :

$$F^{-1}(\pi_i) = \Phi(\gamma^\top \mathbf{W}_i), \tag{2.12}$$

where $\Phi$ is the distribution function of $\mathcal{N}(0, 1)$. According to (2.4)-(2.12) the log-likelihood of $\theta = (\beta^\top, \gamma^\top)^\top$

$$\begin{aligned}
\ell_n^{probit}(\theta) &= \sum_{i=1}^{n} \Big\{ J_i \log\left[\Phi(\gamma^\top \mathbf{W}_i) + (1 - \Phi(\gamma^\top \mathbf{W}_i))e^{-\exp(\beta^\top \mathbf{X}_i)}\right] \\
&\quad + (1 - J_i)\left[Z_i\beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} + \log\left(1 - \Phi(\gamma^\top \mathbf{W}_i)\right) - \log(Z_i!)\right]\Big\}.
\end{aligned}$$

The MLE $\hat{\theta}_n = (\hat{\beta}_n^\top, \hat{\gamma}_n^\top)^\top$ of $\theta$ is the solution of the $k$-dimensional score equation

$$\frac{\partial \ell_n^{probit}(\theta)}{\partial \theta} = 0 \tag{2.13}$$

Solving this (non-linear) equation is relatively straightforward using standard mathematical softwares

**Remark 1.** *A rigorous assessment of the asymptotic properties of $\hat{\theta}_n$ is presented in the censored ZIP model Nguyen et al. (2019). In this paper, such properties can be expected in the ZIP model regardless of the link function used to model the probability of susceptibility. However, leaving aside the distribution theory, we propose to study these properties by means of simulations.*

## 3. A simulation study

In this section, we compare, via simulations, the performance of four links functions (2.4)-(2.7)-(2.10)-(2.12) used to model the probability of zero-inflated. We generate 2 covariates for our simulation study $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{W}_i = (1, W_{i2}, \cdots, W_{iq})^\top$, where $X_{i1} = W_{i1} = 1$ and the $X_{i2}, \cdots, X_{i6}, W_{i4}, W_{i5}$ are independently drawn from normal $\mathcal{N}(0,1)$,binomial $\mathcal{B}(1, 0.3)$, normal $\mathcal{N}(1, 1.5)$, exponential $\mathcal{E}(1)$, uniform $\mathcal{U}(2, 5)$, normal $\mathcal{N}(-1, 1)$ and binomial $\mathcal{B}(1, 0.5)$ distributions respectively. Linear predictors are allowed to share common terms by letting $W_{i2} = X_{i2}$ et $W_{i3} = X_{i3}$. The regression parameter $\beta$ is chosen as $\beta = (0.7, 0.1, 0.4, 0.85, -0.5, 0)^\top$ for all simulations. With the same value of $\beta$, we carry out our studies under four scenarios based on four true models as follows

### *3.1. Simulation scenario*

**Scenario 1**: The following ZIP regression model is used to simulate data :

$$\log(\lambda_i(\beta)) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6},$$

$$\text{logit}(\pi_i) = \gamma_1 W_{i1} + \gamma_2 W_{i2} + \gamma_3 W_{i3} + \gamma_4 W_{i4}, + \gamma_5 W_{i5},$$

The regression parameter $\gamma$ is chosen as $\gamma = (-0.9, -0.65, -0.2, 0.65, 0)^\top$. In this setting, the average proportion of zero-inflated data is 0.20.

**Scenario 2**: The count data are generated from the cloglog link model with $F^{-1}(\pi_i) = -\log(-\log(\pi_i)) = \gamma^\top \mathbf{W}_i$. We consider two values for $\gamma$, namely : $\gamma = (0.5, -0.60, -0.2, 0.75, 0)^\top$ and $\gamma = (0.25, -0.9, 0.60, -0.45, 0)^\top$. With these values, the average proportion $c$ of zero-inflated data in the simulated data sets is 0.20 and 0.60 respectively.

**Scenario 3:** We simulate the data according to the ZIP-GEV model (2.7)-(2.8) defined by:

$$\log(\lambda_i(\beta)) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6},$$

$$\text{gevit}(\omega_i) = \gamma_1 W_{i1} + \gamma_2 W_{i2} + \gamma_3 W_{i3} + \gamma_4 W_{i4}, + \gamma_5 W_{i5},$$

The advantage of the GEV link model we are talking about here is that it integrates a wide range of asymmetries with the shape parameter $\xi$. But in our simulations we choose $\xi = 0.5$ belonging to Frechet's domain. The regression parameter $\gamma$ is chosen as:

- case 1: $\gamma = (-0.95, 0.5, -0.4, -0.65, 0)^\top$

- case 2: $\gamma = (0.8, 0.2, 0.4, -0.8, 0)^\top$.

Using these values, in case 1 (respectively case 2), the average percentage of zero-infation in the simulated data sets is 0.25 (respectively 0.75).

**Scenario 4**: In a second set of simulation scenarios, the data sets simulated from the ZIP-probit model with $F^{-1}(\pi_i) = \Phi^{-1}(\gamma^\top \mathbf{W}_i)$.Two values for $\gamma$, namely : $\gamma = (-0.5, 0, 0.65, 0.8, 0)^\top$ and $\gamma = (-0.1, 0.85, 0.9, -0.2, 0)^\top$.The parameter vector $\gamma \in \mathbb{R}^5$ is chosen to yield various average proportions of zero-inflation within each sample, namely : 0.20 and 0.60.

We consider the following sample sizes: $n = 500; 2000$. For each combination of the simulation design parameters (sample size, proportions of zero-inflation),We simulate $N = 1000$ replications for each combination `[sample size` $\times$ `proportion of zero-inflation]` of the design parameters. Simulations are conducted using the statistical software `R` R Core Team (2018). We use the package `maxLik` Henningsen and Toomet (2011) to solve the score equation (2.9)-(2.11)-(2.13) via a Newton-Raphson algorithm.

*3.2. Results*

For each configuration sample `size`$\times$ `zero-inflation proportion` of the simulation design parameters, we calculate the average bias, standard deviation, average standard error and root mean square error of the estimate over the $N$ simulated samples. We also obtain the empirical coverage probability and average length of 95%-level Wald confidence intervals for the $\beta_j$. The results are described in Table 1 (first scenario), Table 2 and Table 3 (second scenario), Table 4 and Table 5 (third scenario), Table 6 and Table7 (fourth scenario) for $n = 500$ and $n = 2000$ respectively.
Through simulations, we also assess the normal approximation by plotting estimated densities obtained from the $N$ normalized estimates $(\hat{\beta}_{j,n} - \beta_j)/\text{standard error}(\hat{\beta}_{j,n})$ $j, k = 1, \ldots, 6$, and by comparing with the density of the standard normal distribution. Standard errors are obtained as the square roots of the diagonal elements of the estimated variance matrix for ours models. Figures 1, 2 and 3 provide results for ZIP-GEV model ($n = 500$ ,25% of zero-inflation), ZIP-cloglog ($n = 500$ and 20% of zero-inflation), and ZIP-probit model ($n = 500$ ,30% of zero-inflation) . Plots for the other scenarios are similar and thus are not given.
From these results, it appears, as expected, that the bias, variability and length of confidence intervals of all estimates decrease as the sample size increases. For fixed $n$, we observe that performances of the $\hat{\beta}_{j,n}$s remain stable when the proportion of zero-inflation varies from small to moderate values.

These observations illustrate the general fact that accurate estimation in a zeroinflated regression model requires a balance between susceptible and non susceptible subpopulations (that is, a sufficient amount of zero and non zero observations should be available to accurately estimate the zero-inflation probabilities and count submodel). Also, empirical coverage probabilities are close to the nominal level, which indicates that the normal approximation of the distribution of the MLE is appropriate, even when the sample size is moderate. This is confirmed by Figures 1 2 and 3.

*3.3. A comparison of the four models*

In this section, we compare, through simulations, the performance of four models . We obtain the EML in the four models, for the four scenarios. In the first scenario, our real model is the ZIP where the probability of zero inflation is modeled by the logit link. The other models ( ZIP-probit, ZIP-cloglog and ZIP-GEV ) misspecifies the susceptibility probability $\pi_i$. In the second scenario, the true model is the ZIP-cloglog . In the third scenario, the true model is the ZIP-cloglog. In the fourth scenario, the true model is the ZIP-probit. In all four cases, the $\gamma$ estimates are assumed to be biased in the misspecified model. This is confirmed by the simulation results. However, in all four scenarios, the interest is generally on the $\beta$ , which relates the covariates to the $\lambda_i$ intensity of the account response. For this reason, we provide results only for $\beta$. Moreover, since the proposed models adopt the same specification for $\pi_i$, a comparison of the $\beta$'s estimates of the four models is fair. The results are described in Table 1 (Scenario 1), Table 2 and Table 3 (Scenario 2), Table 4 and Table 5 (Scenario 3), Table 6 and Table7 (Scenario 4).

It appears that in all four models, the estimate of $\beta$ is quite robust to a misspecification of the probability of susceptibility. That is, when the logit model is used to generate the data (scenario 1), the $\beta$ estimates in the ZIP-Probit and ZIP-GEV models are of good quality. Referring to scenario 2 and scenario 3 described above, we validate the $\beta$ estimates in the ZIP-probit and ZIP-GEV models.

Conversely, when the ZIP-probit and ZIP-GEV models are used to simulate the data (Scenario 3 and Scenario 4), the $\beta$ estimates in these models perform equally well and better than the others proposed. We also observe that the estimates obtained from the ZIP-probit and ZIP-GEV models behave almost systematically better than the estimates based on the other models, even when the ZIP-logit or ZIP-cloglog is used to simulate the data.

| | | Sample size $n = 500$ | | | | | | Sample size $n = 2000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ |
| bias | **logit** | **-0.0249** | **0.0046** | **0.0043** | **0.0083** | **-0.0072** | **-0.0005** | **-0.0250** | **0.0004** | **0.0038** | **0.0072** | **-0.0055** | **0.0000** |
| | cloglog | 0.0008 | 0.0001 | -0.0001 | 0.0003 | -0.0004 | -0.0004 | -0.0007 | 0.0000 | -0.0001 | 0.0000 | 0.0010 | -0.0001 |
| | probit | 0.0010 | 0.0002 | -0.0001 | 0.0002 | -0.0004 | -0.0004 | -0.0006 | 0.0000 | -0.0001 | 0.0000 | 0.0010 | -0.0001 |
| | GEV | 0.0006 | 0.0002 | 0.0000 | 0.0003 | -0.0004 | -0.0004 | -0.0008 | 0.0001 | -0.0001 | 0.0000 | 0.0010 | -0.0001 |
| SD | **logit** | **0.0879** | **0.0198** | **0.0369** | **0.0134** | **0.0285** | **0.0214** | **0.0438** | **0.0096** | **0.0183** | **0.0063** | **0.0137** | **0.0106** |
| | cloglog | 0.0865 | 0.0194 | 0.0364 | 0.0130 | 0.0277 | 0.0210 | 0.0429 | 0.0095 | 0.0180 | 0.0061 | 0.0134 | 0.0104 |
| | probit | 0.0865 | 0.0194 | 0.0364 | 0.0130 | 0.0277 | 0.0210 | 0.0429 | 0.0095 | 0.0180 | 0.0061 | 0.0134 | 0.0104 |
| | GEV | 0.0866 | 0.0195 | 0.0364 | 0.0130 | 0.0278 | 0.0210 | 0.0429 | 0.0095 | 0.0180 | 0.0061 | 0.0134 | 0.0104 |
| SE | **logit** | **0.0897** | **0.0191** | **0.0379** | **0.0136** | **0.0285** | **0.0214** | **0.0435** | **0.0092** | **0.0183** | **0.0063** | **0.0137** | **0.0104** |
| | cloglog | 0.0882 | 0.0189 | 0.0376 | 0.0133 | 0.0277 | 0.0212 | 0.0428 | 0.0091 | 0.0182 | 0.0062 | 0.0134 | 0.0103 |
| | probit | 0.0882 | 0.0189 | 0.0376 | 0.0133 | 0.0277 | 0.0212 | 0.0428 | 0.0091 | 0.0182 | 0.0062 | 0.0134 | 0.0103 |
| | GEV | 0.0882 | 0.0189 | 0.0376 | 0.0133 | 0.0277 | 0.0212 | 0.0428 | 0.0091 | 0.0182 | 0.0062 | 0.0134 | 0.0103 |
| RMSE | **logit** | **0.1280** | **0.0279** | **0.0531** | **0.0208** | **0.0409** | **0.0302** | **0.0666** | **0.0139** | **0.0262** | **0.0115** | **0.0202** | **0.0148** |
| | cloglog | 0.1236 | 0.0271 | 0.0523 | 0.0185 | 0.0392 | 0.0299 | 0.0606 | 0.0131 | 0.0256 | 0.0087 | 0.0190 | 0.0146 |
| | probit | 0.1235 | 0.0271 | 0.0523 | 0.0185 | 0.0392 | 0.0298 | 0.0606 | 0.0131 | 0.0256 | 0.0087 | 0.0190 | 0.0146 |
| | GEV | 0.1236 | 0.0271 | 0.0523 | 0.0186 | 0.0392 | 0.0299 | 0.0606 | 0.0132 | 0.0256 | 0.0087 | 0.0190 | 0.0146 |
| CP | **logit** | **0.9460** | **0.9330** | **0.9500** | **0.9080** | **0.9430** | **0.9480** | **0.909** | **0.9070** | **0.9440** | **0.7970** | **0.9300** | **0.9340** |
| | cloglog | 0.9510 | 0.9450 | 0.9530 | 0.9590 | 0.9450 | 0.9450 | 0.9570 | 0.9490 | 0.9530 | 0.9580 | 0.9450 | 0.937 |
| | probit | 0.9510 | 0.9430 | 0.9530 | 0.9610 | 0.9450 | 0.9460 | 0.9570 | 0.9500 | 0.9540 | 0.9580 | 0.9450 | 0.9370 |
| | GEV | 0.9490 | 0.9460 | 0.9510 | 0.9590 | 0.9450 | 0.9440 | 0.9570 | 0.94800 | 0.9530 | 0.9570 | 0.9450 | 0.9370 |
| $\ell$(CI) | **logit** | **0.3504** | **0.0746** | **0.1482** | **0.0527** | **0.1111** | **0.0838** | **0.1702** | **0.0361** | **0.0717** | **0.0247** | **0.0538** | **0.0406** |
| | cloglog | 0.3449 | 0.0739 | 0.1470 | 0.0515 | 0.1083 | 0.0828 | 0.1678 | 0.0358 | 0.0712 | 0.0242 | 0.0525 | 0.0402 |
| | probit | 0.3449 | 0.0739 | 0.1470 | 0.0515 | 0.1083 | 0.0828 | 0.1678 | 0.0358 | 0.0712 | 0.0242 | 0.0525 | 0.0402 |
| | GEV | 0.3450 | 0.0739 | 0.1470 | 0.0515 | 0.1083 | 0.0828 | 0.1678 | 0.0358 | 0.0712 | 0.0242 | 0.0525 | 0.0402 |

Table 1: Simulation results (data are simulated from the ZIP model (2.6)-(2.3)-(2.4), ZI proportion = 20%).
SD: empirical standard deviation. SE: average standard error. CP: empirical coverage probability of 95%-level confidence intervals. $\ell$(CI): average length of confidence intervals.

| | | Sample size $n = 500$ | | | | | | Sample size $n = 2000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ |
| bias | **cloglog** | **-0.0031** | **0.0007** | **-0.0015** | **-0.0001** | **-0.0008** | **0.0009** | **0.0003** | **-0.0004** | **-0.0003** | **0.0000** | **0.0003** | **0 .0000** |
| | GEV | -0.0036 | 0.0008 | -0.0014 | 0.0000 | -0.0009 | 0.0009 | -0.0001 | -0.0004 | -0.0003 | 0.0001 | 0.0002 | 0.0000 |
| | logit | -0.0503 | 0.0096 | 0.0065 | 0.0141 | -0.0130 | 0.0001 | -0.0434 | 0.0076 | 0.0072 | 0.0126 | -0.0110 | 0.0000 |
| | probit | -0.0024 | 0.0011 | -0.0015 | -0.0004 | -0.0007 | 0.0009 | 0.0010 | -0.0001 | -0.0003 | -0.0002 | 0.0004 | 0.0000 |
| SD | **cloglog** | **0.0983** | **0.0203** | **0.0382** | **0.0141** | **0.0300** | **0.0236** | **0.0436** | **0.0099** | **0.0199** | **0.0066** | **0.0144** | **0.0104** |
| | GEV | 0.0985 | 0.0204 | 0.0385 | 0.0142 | 0.0300 | 0.0236 | 0.0436 | 0.0099 | 0.0199 | 0.0066 | 0.0144 | 0.0104 |
| | logit | 0.1028 | 0.0210 | 0.0390 | 0.0149 | 0.0314 | 0.0243 | 0.0455 | 0.0102 | 0.0202 | 0.0071 | 0.0152 | 0.0108 |
| | probit | 0.0983 | 0.0203 | 0.0382 | 0.0141 | 0.0300 | 0.0236 | 0.0436 | 0.0099 | 0.0199 | 0.0066 | 0.0144 | 0.0104 |
| SE | **cloglog** | **0.0944** | **0.0206** | **0.0399** | **0.0142** | **0.0296** | **0.0226** | **0.0456** | **0.0099** | **0.0193** | **0.0066** | **0.0143** | **0.0109** |
| | GEV | 0.0944 | 0.0206 | 0.0399 | 0.0142 | 0.0296 | 0.0226 | 0.0455 | 0.0099 | 0.0193 | 0.0066 | 0.0143 | 0.0109 |
| | logit | 0.0965 | 0.0209 | 0.0404 | 0.0147 | 0.0305 | 0.0230 | 0.0464 | 0.0100 | 0.0195 | 0.0068 | 0.0147 | 0.0111 |
| | probit | 0.0945 | 0.0206 | 0.0399 | 0.0143 | 0.0296 | 0.0226 | 0.0456 | 0.0099 | 0.0193 | 0.0066 | 0.0143 | 0.0109 |
| RMSE | **cloglog** | **0.1363** | **0.0289** | **0.0552** | **0.0200** | **0.0421** | **0.0327** | **0.0630** | **0.0140** | **0.0277** | **0.0093** | **0.0203** | **0.0151** |
| | GEV | 0.1364 | 0.0289 | 0.0554 | 0.0201 | 0.0421 | 0.0327 | 0.0630 | 0.0140 | 0.0277 | 0.0093 | 0.0203 | 0.0151 |
| | logit | 0.1497 | 0.0311 | 0.0565 | 0.0252 | 0.0457 | 0.0335 | 0.0782 | 0.0162 | 0.0290 | 0.0160 | 0.0238 | 0.0154 |
| | probit | 0.1363 | 0.0289 | 0.0552 | 0.0201 | 0.0421 | 0.0327 | 0.0630 | 0.0140 | 0.0277 | 0.0093 | 0.0203 | 0.0150 |
| CP | **cloglog** | **0.9340** | **0.9480** | **0.9610** | **0.9520** | **0.9440** | **0.9340** | **0.9610** | **0.9500** | **0.9470** | **0.9490** | **0.9480** | **0.9620** |
| | GEV | 0.9320 | 0.9480 | 0.9590 | 0.9510 | 0.9440 | 0.9350 | 0.9600 | 0.9470 | 0.9450 | 0.9510 | 0.9490 | 0.9620 |
| | logit | 0.9140 | 0.9160 | 0.9530 | 0.8310 | 0.9260 | 0.9300 | 0.8460 | 0.8680 | 0.9270 | 0.5290 | 0.8840 | 0.9590 |
| | probit | 0.9360 | 0.9480 | 0.9590 | 0.9520 | 0.9420 | 0.9350 | 0.9620 | 0.9530 | 0.9450 | 0.9500 | 0.9500 | 0.9620 |
| $\ell$(CI) | **cloglog** | **0.3691** | **0.0803** | **0.1560** | **0.0552** | **0.1154** | **0.0884** | **0.1784** | **0.0388** | **0.0756** | **0.0258** | **0.0559** | **0.0427** |
| | GEV | 0.3691 | 0.0803 | 0.1560 | 0.0552 | 0.1154 | 0.0884 | 0.1784 | 0.0388 | 0.0756 | 0.0258 | 0.0559 | 0.0427 |
| | logit | 0.3771 | 0.0815 | 0.1579 | 0.0569 | 0.1192 | 0.0898 | 0.1818 | 0.0392 | 0.0763 | 0.0265 | 0.0576 | 0.0433 |
| | probit | 0.3694 | 0.0803 | 0.1560 | 0.0553 | 0.1155 | 0.0885 | 0.1785 | 0.0388 | 0.0756 | 0.0259 | 0.0560 | 0.0427 |

Table 2: Simulation results (data are simulated from the ZIP-cloglog model (2.6)-(2.4)-(2.10), ZI proportion = 30%). SD: empirical standard deviation. SE: average standard error. CP: empirical coverage probability of 95%-level confidence intervals. $\ell$(CI): average length of confidence intervals.

| | | Sample size $n = 500$ | | | | | | Sample size $n = 2000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ |
| bias | **cloglog** | **-0.0055** | **-0.0023** | **0.0022** | **0.0008** | **-0.0024** | **0.0013** | **0.0009** | **-0.0006** | **0.0001** | **0.0001** | **-0.0007** | **-0.0001** |
| | GEV | -0.0069 | -0.0023 | 0.0024 | 0.0012 | -0.0027 | 0.0013 | -0.0001 | -0.0005 | 0.0007 | 0.0003 | -0.0008 | -0.0001 |
| | logit | -0.2022 | 0.0472 | 0.0010 | 0.0575 | -0.0487 | 0.0020 | -0.1692 | 0.0416 | -0.0002 | 0.0475 | -0.0425 | -0.0002 |
| | probit | -0.0047 | -0.0014 | 0.0015 | 0.0004 | -0.0022 | 0.0013 | 0.0016 | 0.0003 | 0.0004 | -0.0002 | -0.0004 | -0.0001 |
| SD | **cloglog** | **0.1306** | **0.0293** | **0.0587** | **0.0200** | **0.0410** | **0.0320** | **0.0621** | **0.0138** | **0.0271** | **0.0093** | **0.0187** | **0.0147** |
| | GEV | 0.1323 | 0.0294 | 0.0591 | 0.0205 | 0.0411 | 0.0320 | 0.0623 | 0.0138 | 0.0286 | 0.0093 | 0.0187 | 0.0147 |
| | logit | 0.1599 | 0.0381 | 0.0686 | 0.0281 | 0.0489 | 0.0372 | 0.0767 | 0.0177 | 0.0327 | 0.0139 | 0.0226 | 0.0174 |
| | probit | 0.1304 | 0.0293 | 0.0587 | 0.0200 | 0.0410 | 0.0320 | 0.0620 | 0.0138 | 0.0271 | 0.0093 | 0.0187 | 0.0147 |
| SE | **cloglog** | **0.1308** | **0.0300** | **0.0593** | **0.0203** | **0.0407** | **0.0312** | **0.0619** | **0.0141** | **0.0278** | **0.0091** | **0.0194** | **0.0148** |
| | GEV | 0.1308 | 0.0300 | 0.0593 | 0.0203 | 0.0406 | 0.0312 | 0.0619 | 0.0141 | 0.0278 | 0.0091 | 0.0194 | 0.0148 |
| | logit | 0.1371 | 0.0314 | 0.0615 | 0.0212 | 0.0430 | 0.0324 | 0.0643 | 0.0146 | 0.0286 | 0.0094 | 0.0204 | 0.0152 |
| | probit | 0.1309 | 0.0300 | 0.0593 | 0.0204 | 0.0407 | 0.0312 | 0.0619 | 0.0141 | 0.0278 | 0.0092 | 0.0194 | 0.0148 |
| RMSE | **cloglog** | **0.1849** | **0.0420** | **0.0834** | **0.0286** | **0.0578** | **0.0447** | **0.0876** | **0.0197** | **0.0389** | **0.0130** | **0.0270** | **0.0208** |
| | GEV | 0.1861 | 0.0421 | 0.0837 | 0.0289 | 0.0578 | 0.0447 | 0.0878 | 0.0197 | 0.0399 | 0.0131 | 0.0270 | 0.0208 |
| | logit | 0.2919 | 0.0683 | 0.0921 | 0.0674 | 0.0813 | 0.0494 | 0.1966 | 0.0475 | 0.0434 | 0.0503 | 0.0522 | 0.0231 |
| | probit | 0.1848 | 0.0419 | 0.0834 | 0.0286 | 0.0578 | 0.0447 | 0.0876 | 0.0197 | 0.0389 | 0.0130 | 0.0270 | 0.0208 |
| CP | **cloglog** | **0.9520** | **0.9480** | **0.9530** | **0.9470** | **0.9520** | **0.9470** | **0.9470** | **0.9580** | **0.9570** | **0.9520** | **0.9610** | **0.9620** |
| | GEV | 0.9530 | 0.9500 | 0.9480 | 0.9470 | 0.9520 | 0.9480 | 0.9460 | 0.9530 | 0.9550 | 0.9500 | 0.9620 | 0.9620 |
| | Logit | 0.6630 | 0.6640 | 0.9210 | 0.2500 | 0.7830 | 0.9020 | 0.2700 | 0.2290 | 0.9130 | 0.0060 | 0.4530 | 0.9140 |
| | Probit | 0.9560 | 0.9500 | 0.9530 | 0.9480 | 0.9510 | 0.9480 | 0.9480 | 0.9560 | 0.9590 | 0.9510 | 0.9620 | 0.9590 |
| $\ell$(CI) | **cloglog** | **0.5102** | **0.1166** | **0.2309** | **0.0786** | **0.1582** | **0.1217** | **0.2421** | **0.0551** | **0.1089** | **0.0356** | **0.0760** | **0.0578** |
| | GEV | 0.5101 | 0.1167 | 0.2309 | 0.0785 | 0.1581 | 0.1217 | 0.2421 | 0.0552 | 0.1089 | 0.0356 | 0.0760 | 0.0578 |
| | Logit | 0.5344 | 0.1221 | 0.2391 | 0.0817 | 0.1672 | 0.1262 | 0.2516 | 0.0572 | 0.1120 | 0.0366 | 0.0799 | 0.0595 |
| | Probit | 0.5105 | 0.1166 | 0.2309 | 0.0787 | 0.1583 | 0.1218 | 0.2423 | 0.0551 | 0.1089 | 0.0357 | 0.0761 | 0.0578 |

Table 3: Simulation results (data are simulated from the ZIP-cloglog model (2.6)-(2.4)-(2.10), ZI proportion = 60%). SD: empirical standard deviation. SE: average standard error. CP: empirical coverage probability of 95%-level confidence intervals. $\ell$(CI): average length of confidence intervals.

| | | Sample size $n = 500$ | | | | | | Sample size $n = 2000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ |
| bias | **GEV** | **-0.0009** | **-0.0012** | **-0.0000** | **-0.0003** | **0.0004** | **0.0002** | **-0.0014** | **0.0000** | **-0.0002** | **0.0000** | **0.0002** | **0.0003** |
| | logit | -0.0346 | -0.0026 | 0.0078 | 0.0102 | -0.0086 | 0.0001 | -0.0324 | -0.0013 | 0.0066 | 0.0093 | -0.0083 | 0.0004 |
| | cloglog | -0.0004 | -0.0012 | -0.0000 | -0.0004 | 0.0005 | 0.0002 | -0.0011 | -0.0001 | -0.0002 | 0.0000 | 0.0002 | 0.0003 |
| | probit | -0.0003 | -0.0014 | 0.0002 | -0.0005 | 0.0006 | 0.0002 | -0.0006 | -0.0003 | 0.0000 | -0.0002 | 0.0003 | 0.0003 |
| SD | **GEV** | **0.0902** | **0.0193** | **0.0373** | **0.0138** | **0.0283** | **0.0223** | **0.0439** | **0.0092** | **0.0183** | **0.0063** | **0.0139** | **0.0107** |
| | logit | 0.0917 | 0.0196 | 0.0382 | 0.0143 | 0.0294 | 0.0227 | 0.0449 | 0.0094 | 0.0187 | 0.0067 | 0.0143 | 0.0109 |
| | cloglog | 0.0902 | 0.0193 | 0.0373 | 0.0138 | 0.0283 | 0.0223 | 0.0439 | 0.0092 | 0.0183 | 0.0063 | 0.0139 | 0.0107 |
| | probit | 0.0902 | 0.0192 | 0.0373 | 0.0138 | 0.0283 | 0.0223 | 0.0439 | 0.0092 | 0.0183 | 0.0064 | 0.0139 | 0.0107 |
| SE | **GEV** | **0.0911** | **0.0193** | **0.0384** | **0.0137** | **0.0286** | **0.0219** | **0.0441** | **0.0093** | **0.0186** | **0.0064** | **0.0139** | **0.0106** |
| | logit | 0.0929 | 0.0196 | 0.0388 | 0.0141 | 0.0294 | 0.0222 | 0.0449 | 0.0094 | 0.0187 | 0.0065 | 0.0142 | 0.0107 |
| | cloglog | 0.0911 | 0.0193 | 0.0384 | 0.0137 | 0.0286 | 0.0219 | 0.0442 | 0.0093 | 0.0186 | 0.0064 | 0.0139 | 0.0106 |
| | probit | 0.0912 | 0.0193 | 0.0384 | 0.0137 | 0.0286 | 0.0219 | 0.0442 | 0.0093 | 0.0186 | 0.0064 | 0.0139 | 0.0106 |
| RMSE | **GEV** | **0.1282** | **0.0273** | **0.0535** | **0.0195** | **0.0402** | **0.0312** | **0.0623** | **0.0131** | **0.0261** | **0.0090** | **0.0196** | **0.0151** |
| | logit | 0.1350 | 0.0278 | 0.0550 | 0.0226 | 0.0425 | 0.0318 | 0.0713 | 0.0133 | 0.0273 | 0.0132 | 0.0218 | 0.0153 |
| | cloglog | 0.1282 | 0.0273 | 0.0535 | 0.0195 | 0.0402 | 0.0312 | 0.0623 | 0.0131 | 0.0260 | 0.0090 | 0.0196 | 0.0151 |
| | probit | 0.1282 | 0.0273 | 0.0535 | 0.0195 | 0.0402 | 0.0313 | 0.0622 | 0.0131 | 0.0260 | 0.0090 | 0.0196 | 0.0151 |
| CP | **GEV** | **0.9440** | **0.9530** | **0.9610** | **0.9440** | **0.9470** | **0.9460** | **0.9450** | **0.9530** | **0.9580** | **0.9540** | **0.9500** | **0.9590** |
| | logit | 0.9370 | 0.9470 | 0.9580 | 0.8860 | 0.9420 | 0.9470 | 0.9020 | 0.9470 | 0.9400 | 0.7080 | 0.9170 | 0.9490 |
| | cloglog | 0.9440 | 0.9550 | 0.9620 | 0.9430 | 0.9460 | 0.9460 | 0.9440 | 0.9520 | 0.9580 | 0.9540 | 0.9500 | 0.9600 |
| | probit | 0.9450 | 0.9560 | 0.9620 | 0.9430 | 0.9440 | 0.9450 | 0.9450 | 0.9530 | 0.9580 | 0.9520 | 0.9500 | 0.9600 |
| $\ell$(CI) | **GEV** | **0.3560** | **0.0754** | **0.1503** | **0.0532** | **0.1115** | **0.0857** | **0.1729** | **0.0363** | **0.0727** | **0.0249** | **0.0542** | **0.0415** |
| | logit | 0.3628 | 0.0763 | 0.1519 | 0.0548 | 0.1149 | 0.0868 | 0.1759 | 0.0367 | 0.0734 | 0.0255 | 0.0558 | 0.0420 |
| | cloglog | 0.3561 | 0.0754 | 0.1503 | 0.0533 | 0.1116 | 0.0857 | 0.1729 | 0.0363 | 0.0727 | 0.0249 | 0.0543 | 0.0415 |
| | probit | 0.3562 | 0.0754 | 0.1503 | 0.0533 | 0.1117 | 0.0857 | 0.1730 | 0.0363 | 0.0727 | 0.0249 | 0.0543 | 0.0415 |

Table 4: Simulation results (data are simulated from the ZIP-GEV model (2.7)-(2.8), ZI proportion = 25%). SD: empirical standard deviation. SE: average standard error. CP: empirical coverage probability of 95%-level confidence intervals. $\ell$(CI): average length of confidence intervals.

| | | Sample size $n = 500$ | | | | | | Sample size $n = 2000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ |
| bias | **GEV** | **-0.0123** | **-0.0021** | **-0.0009** | **0.0013** | **-0.0019** | **0.0022** | **0.0005** | **0.0002** | **-0.0015** | **-0.0004** | **-0.0009** | **0.0000** |
| | logit | -0.3718 | -0.0022 | 0.0291 | 0.1256 | -0.1053 | 0.0022 | -0.3075 | 0.0008 | 0.0238 | 0.0989 | -0.0864 | -0.0007 |
| | cloglog | -0.0105 | -0.0018 | -0.0008 | 0.0009 | -0.0015 | 0.0021 | 0.0019 | 0.0005 | 0.0004 | -0.0007 | -0.0007 | 0.0002 |
| | probit | -0.0101 | -0.0019 | -0.0010 | 0.0008 | -0.0014 | 0.0021 | 0.0023 | 0.0004 | 0.0002 | -0.0008 | -0.0006 | 0.0002 |
| SD | **GEV** | **0.1685** | **0.0364** | **0.0788** | **0.0273** | **0.0541** | **0.0416** | **0.0814** | **0.0182** | **0.0559** | **0.0129** | **0.0261** | **0.0191** |
| | logit | 0.2475 | 0.0496 | 0.1160 | 0.0503 | 0.0780 | 0.0590 | 0.1181 | 0.0231 | 0.0533 | 0.0250 | 0.0360 | 0.0274 |
| | cloglog | 0.1664 | 0.0355 | 0.0781 | 0.0268 | 0.0534 | 0.0416 | 0.0780 | 0.0160 | 0.0357 | 0.0118 | 0.0253 | 0.0189 |
| | probit | 0.1665 | 0.0355 | 0.0781 | 0.0268 | 0.0534 | 0.0416 | 0.0780 | 0.0160 | 0.0357 | 0.0118 | 0.0253 | 0.0189 |
| SE | **GEV** | **0.1713** | **0.0361** | **0.0767** | **0.0277** | **0.0543** | **0.0412** | **0.0788** | **0.0166** | **0.0349** | **0.0120** | **0.0250** | **0.019** |
| | logit | 0.1860 | 0.0388 | 0.0821 | 0.0293 | 0.0596 | 0.0443 | 0.0841 | 0.0175 | 0.0367 | 0.0125 | 0.0272 | 0.020 |
| | cloglog | 0.1712 | 0.0361 | 0.0767 | 0.0277 | 0.0543 | 0.0412 | 0.0789 | 0.0166 | 0.0349 | 0.0120 | 0.0250 | 0.019 |
| | probit | 0.1712 | 0.0361 | 0.0767 | 0.0277 | 0.0544 | 0.0412 | 0.0788 | 0.0166 | 0.0349 | 0.0120 | 0.0250 | 0.019 |
| RMSE | **GEV** | **0.2405** | **0.0513** | **0.1100** | **0.0389** | **0.0767** | **0.0586** | **0.1133** | **0.0246** | **0.0659** | **0.0176** | **0.0361** | **0.0269** |
| | logit | 0.4837 | 0.0630 | 0.1451 | 0.1384 | 0.1440 | 0.0738 | 0.3400 | 0.0290 | 0.0690 | 0.1028 | 0.0974 | 0.0339 |
| | cloglog | 0.2390 | 0.0507 | 0.1095 | 0.0386 | 0.0762 | 0.0586 | 0.1109 | 0.0230 | 0.0499 | 0.0168 | 0.0356 | 0.0268 |
| | probit | 0.2390 | 0.0507 | 0.1095 | 0.0386 | 0.0762 | 0.0586 | 0.1109 | 0.0230 | 0.0499 | 0.0168 | 0.0356 | 0.0268 |
| CP | **GEV** | **0.9500** | **0.9550** | **0.9440** | **0.9580** | **0.9490** | **0.9510** | **0.9420** | **0.9560** | **0.9420** | **0.9520** | **0.9390** | **0.9470** |
| | logit | 0.4790 | 0.8780 | 0.8180 | 0.0240 | 0.5680 | 0.8570 | 0.0910 | 0.8590 | 0.7870 | 0.0000 | 0.1640 | 0.854 |
| | cloglog | 0.9520 | 0.9580 | 0.9450 | 0.9600 | 0.9500 | 0.9510 | 0.9480 | 0.9610 | 0.9490 | 0.9560 | 0.9440 | 0.9510 |
| | probit | 0.9540 | 0.9580 | 0.9440 | 0.9600 | 0.9500 | 0.9520 | 0.9480 | 0.9610 | 0.9500 | 0.9560 | 0.9430 | 0.9510 |
| $\ell$(CI) | **GEV** | **0.6661** | **0.1400** | **0.2980** | **0.1066** | **0.2107** | **0.1604** | **0.3083** | **0.0647** | **0.1366** | **0.0466** | **0.0978** | **0.0742** |
| | logit | 0.7215 | 0.1500 | 0.3182 | 0.1129 | 0.2307 | 0.1719 | 0.3288 | 0.0682 | 0.1435 | 0.0485 | 0.1060 | 0.0782 |
| | cloglog | 0.6660 | 0.1400 | 0.2980 | 0.1066 | 0.2108 | 0.1604 | 0.3084 | 0.0647 | 0.1366 | 0.0466 | 0.0979 | 0.0742 |
| | probit | 0.6659 | 0.1400 | 0.2980 | 0.1066 | 0.2108 | 0.1604 | 0.3083 | 0.0647 | 0.1366 | 0.0466 | 0.0979 | 0.0742 |

Table 5: Simulation results (data are simulated from the ZIP-GEV model (2.7)-(2.8), ZI proportion = 75%). SD: empirical standard deviation. SE: average standard error. CP: empirical coverage probability of 95%-level confidence intervals. $\ell$(CI): average length of confidence intervals.

| | | Sample size $n = 500$ | | | | | | Sample size $n = 2000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ |
| bias | **probit** | **-0.0024** | **-0.0006** | **0.0013** | **0.0001** | **0.0012** | **0.0002** | **0.0014** | **0.0002** | **-0.0006** | **-0.0003** | **0.0001** | **-0.0001** |
| | logit | -0.0252 | 0.0004 | 0.0004 | 0.0081 | -0.0059 | 0.0001 | -0.0202 | 0.0010 | -0.0015 | 0.0069 | -0.0064 | -0.0001 |
| | GEV | -0.0037 | -0.0006 | 0.0015 | 0.0004 | 0.0011 | 0.0002 | 0.0006 | 0.0002 | -0.0004 | -0.0001 | 0.0001 | -0.0001 |
| | cloglog | -0.0032 | -0.0006 | 0.0017 | 0.0003 | 0.0011 | 0.0002 | 0.0008 | 0.0002 | -0.0004 | -0.0001 | 0.0001 | -0.0001 |
| SD | **probit** | **0.0882** | **0.0196** | **0.0396** | **0.0138** | **0.0273** | **0.0213** | **0.0429** | **0.0088** | **0.0195** | **0.0063** | **0.0135** | **0.0103** |
| | logit | 0.0904 | 0.0198 | 0.0403 | 0.0142 | 0.0282 | 0.0217 | 0.0435 | 0.0089 | 0.0198 | 0.0065 | 0.0139 | 0.0105 |
| | GEV | 0.0886 | 0.0196 | 0.0399 | 0.0139 | 0.0273 | 0.0213 | 0.0430 | 0.0088 | 0.0195 | 0.0063 | 0.0135 | 0.0102 |
| | cloglog | 0.0882 | 0.0196 | 0.0397 | 0.0138 | 0.0273 | 0.0213 | 0.0429 | 0.0088 | 0.0195 | 0.0063 | 0.0135 | 0.0102 |
| SE | **probit** | **0.0888** | **0.0188** | **0.0394** | **0.0134** | **0.0279** | **0.0215** | **0.0432** | **0.0090** | **0.0190** | **0.0063** | **0.0136** | **0.0104** |
| | logit | 0.0905 | 0.0191 | 0.0398 | 0.0139 | 0.0287 | 0.0218 | 0.0439 | 0.0091 | 0.0192 | 0.0064 | 0.0140 | 0.0105 |
| | GEV | 0.0887 | 0.0188 | 0.0394 | 0.0134 | 0.0279 | 0.0215 | 0.0431 | 0.0090 | 0.0190 | 0.0063 | 0.0136 | 0.0104 |
| | cloglog | 0.0887 | 0.0188 | 0.0394 | 0.0134 | 0.0279 | 0.0215 | 0.0431 | 0.0090 | 0.0190 | 0.0063 | 0.0136 | 0.0104 |
| RMSE | **probit** | **0.1252** | **0.0272** | **0.0559** | **0.0192** | **0.0390** | **0.0302** | **0.0609** | **0.0126** | **0.0272** | **0.0089** | **0.0192** | **0.0146** |
| | logit | 0.1303 | 0.0274 | 0.0566 | 0.0214 | 0.0406 | 0.0307 | 0.0650 | 0.0128 | 0.0276 | 0.0115 | 0.0207 | 0.0148 |
| | GEV | 0.1254 | 0.0272 | 0.0561 | 0.0193 | 0.0390 | 0.0302 | 0.0609 | 0.0126 | 0.0272 | 0.0089 | 0.0192 | 0.0146 |
| | cloglog | 0.1252 | 0.0272 | 0.0559 | 0.0192 | 0.0390 | 0.0302 | 0.0609 | 0.0126 | 0.0272 | 0.0089 | 0.0192 | 0.0146 |
| CP | **probit** | **0.9490** | **0.9320** | **0.9530** | **0.9440** | **0.9480** | **0.9560** | **0.9580** | **0.9640** | **0.9430** | **0.9460** | **0.9530** | **0.9580** |
| | logit | 0.9420 | 0.9320 | 0.9490 | 0.9000 | 0.9450 | 0.9550 | 0.9310 | 0.9610 | 0.9410 | 0.8050 | 0.9280 | 0.9520 |
| | GEV | 0.9480 | 0.9320 | 0.9520 | 0.9420 | 0.9480 | 0.9560 | 0.9560 | 0.9630 | 0.9410 | 0.9470 | 0.9540 | 0.9580 |
| | cloglog | 0.9490 | 0.9320 | 0.9530 | 0.9450 | 0.9480 | 0.9570 | 0.9550 | 0.9640 | 0.9430 | 0.9470 | 0.9540 | 0.9580 |
| $\ell$(CI) | **probit** | **0.3472** | **0.0735** | **0.1539** | **0.0522** | **0.1089** | **0.0839** | **0.1690** | **0.0354** | **0.0745** | **0.0244** | **0.0533** | **0.0407** |
| | logit | 0.3539 | 0.0744 | 0.1555 | 0.0538 | 0.1122 | 0.0851 | 0.1720 | 0.0357 | 0.0752 | 0.0251 | 0.0548 | 0.0412 |
| | GEV | 0.3470 | 0.0735 | 0.1539 | 0.0521 | 0.1089 | 0.0839 | 0.1689 | 0.0354 | 0.0745 | 0.0244 | 0.0532 | 0.0407 |
| | cloglog | 0.3471 | 0.0736 | 0.1539 | 0.0521 | 0.1089 | 0.0839 | 0.1690 | 0.0354 | 0.0745 | 0.0244 | 0.0532 | 0.0407 |

Table 6: Simulation results (data are simulated from the ZIP-Probit model (2.6)-(2.4)-(2.12), ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. CP: empirical coverage probability of 95%-level confidence intervals. $\ell$(CI): average length of confidence intervals.

| | | Sample size $n = 500$ | | | | | | Sample size $n = 2000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ | $\hat{\beta}_{1,n}$ | $\hat{\beta}_{2,n}$ | $\hat{\beta}_{3,n}$ | $\hat{\beta}_{4,n}$ | $\hat{\beta}_{5,n}$ | $\hat{\beta}_{6,n}$ |
| bias | **probit** | **-0.0003** | **-0.0014** | **0.0019** | **0.0004** | **-0.0022** | **-0.0004** | **0.0009** | **0.0001** | **-0.0012** | **-0.0001** | **-0.0013** | **0.0001** |
| | logit | -0.2035 | -0.0603 | -0.0426 | 0.0649 | -0.0576 | -0.0018 | -0.1848 | -0.0489 | -0.0374 | 0.0537 | -0.0494 | 0.0001 |
| | GEV | -0.0033 | 0.0002 | 0.0034 | 0.0013 | -0.0027 | -0.0002 | -0.0012 | 0.0018 | 0.0003 | 0.0005 | -0.0015 | 0.0001 |
| | cloglog | -0.0017 | 0.0001 | 0.0034 | 0.0009 | -0.0025 | -0.0004 | -0.0003 | 0.0015 | 0.0001 | 0.0003 | -0.0015 | 0.0001 |
| SD | **probit** | **0.1326** | **0.0340** | **0.0730** | 0.0214 | **0.0428** | **0.0329** | **0.0624** | **0.0155** | **0.0313** | **0.0099** | **0.0214** | **0.0156** |
| | logit | 0.1664 | 0.0455 | 0.0915 | 0.0320 | 0.0524 | 0.0404 | 0.0796 | 0.0205 | 0.0400 | 0.0154 | 0.0264 | 0.0185 |
| | GEV | 0.1334 | 0.0343 | 0.0730 | 0.0216 | 0.0431 | 0.0331 | 0.0625 | 0.0158 | 0.0314 | 0.0100 | 0.0214 | 0.0157 |
| | ploglog | 0.1327 | 0.0341 | 0.0729 | 0.0215 | 0.0429 | 0.0330 | 0.0624 | 0.0156 | 0.0313 | 0.0100 | 0.0214 | 0.0156 |
| SE | **probit** | **0.1393** | **0.0347** | **0.0723** | **0.0220** | **0.0442** | **0.0336** | **0.066** | **0.0162** | **0.0332** | **0.0098** | **0.0208** | **0.0158** |
| | logit | 0.1468 | 0.0368 | 0.0769 | 0.0229 | 0.0472 | 0.0350 | 0.069 | 0.0170 | 0.0347 | 0.0102 | 0.0221 | 0.0163 |
| | GEV | 0.1392 | 0.0348 | 0.0724 | 0.0219 | 0.0442 | 0.0336 | 0.066 | 0.0162 | 0.0332 | 0.0098 | 0.0208 | 0.0158 |
| | cloglog | 0.1393 | 0.0348 | 0.0724 | 0.0219 | 0.0442 | 0.0336 | 0.066 | 0.0162 | 0.0332 | 0.0098 | 0.0208 | 0.0158 |
| RMSE | **probit** | **0.1923** | **0.0486** | **0.1027** | **0.0307** | **0.0615** | **0.0470** | **0.0909** | **0.0224** | **0.0456** | **0.0140** | **0.0299** | **0.0222** |
| | logit | 0.3011 | 0.0840 | 0.1269 | 0.0759 | 0.0910 | 0.0535 | 0.2127 | 0.0556 | 0.0648 | 0.0567 | 0.0602 | 0.0247 |
| | GEV | 0.1928 | 0.0488 | 0.1028 | 0.0308 | 0.0618 | 0.0471 | 0.0909 | 0.0227 | 0.0457 | 0.0140 | 0.0299 | 0.0222 |
| | cloglog | 0.1923 | 0.0487 | 0.1028 | 0.0307 | 0.0616 | 0.0470 | 0.0909 | 0.0225 | 0.0456 | 0.0140 | 0.0299 | 0.0222 |
| CP | **probit** | **0.9580** | **0.9530** | **0.9550** | **0.9600** | **0.9570** | **0.9520** | **0.9680** | **0.9580** | **0.9620** | **0.9530** | **0.9450** | **0.9480** |
| | logit | 0.7000 | 0.6180 | 0.8750 | 0.2250 | 0.7580 | 0.9120 | 0.2680 | 0.2280 | 0.7770 | 0.0000 | 0.4160 | 0.9160 |
| | GEV | 0.9580 | 0.9470 | 0.9550 | 0.9560 | 0.9560 | 0.9550 | 0.9710 | 0.9540 | 0.9630 | 0.9500 | 0.9460 | 0.9480 |
| | cloglog | 0.9580 | 0.9490 | 0.9540 | 0.9580 | 0.9570 | 0.9540 | 0.9680 | 0.9540 | 0.9620 | 0.9480 | 0.9480 | 0.9480 |
| $\ell$(CI) | **probit** | **0.5433** | **0.1352** | **0.2804** | **0.0849** | **0.1719** | **0.1310** | **0.2585** | **0.0635** | **0.1297** | **0.0383** | **0.0815** | **0.0619** |
| | logit | 0.5719 | 0.1429 | 0.2969 | 0.0886 | 0.1832 | 0.1365 | 0.2700 | 0.0664 | 0.1355 | 0.0396 | 0.0863 | 0.0639 |
| | GEV | 0.5430 | 0.1353 | 0.2805 | 0.0847 | 0.1718 | 0.130 | 0.2584 | 0.0635 | 0.1298 | 0.0383 | 0.0814 | 0.0619 |
| | cloglog | 0.5431 | 0.1353 | 0.2804 | 0.0848 | 0.1718 | 0.1310 | 0.2584 | 0.0635 | 0.1298 | 0.0383 | 0.0814 | 0.0619 |

Table 7: Simulation results (data are simulated from the ZIP-Probit model (2.6)-(2.4)-(2.12), ZI proportion = 60%). SD: empirical standard deviation. SE: average standard error. CP: empirical coverage probability of 95%-level confidence intervals. $\ell$(CI): average length of confidence intervals.

## 4. Applications with real-life data

### *4.1. Data description and competing models*

The data are obtained from the National Medical Expenditure Survey (NMES) which was conducted in 1987 and 1988 to provide a comprehensive picture of how Americans use and pay for health services. The NMES is based upon a representative, national probability sample of the civilian, non-institutionalized population and individuals admitted to long-term care facilities during 1987. Under the household survey of the NMES, more than 38000 individuals in 15000 households across the United States were interviewed quarterly about their health insurance coverage, the services they used, and the cost and source of payments of those services. These data were verified by cross-checking information provided by survey respondents with providers of health-care services. In addition to health-care data, NMES provides information on health status, employment, sociodemographic characteristics, and economic status.

In this paper we consider a subsample of individuals ages 66 and over (a total of 4406 observations) all of whom are covered by Medicare, a public insurance programme that offers substantial protection against health care costs. Residents of the United States are eligible for Medicare coverage at age 65. Some individuals start receiving Medicare benefits a few months into their 65th year primarily because they fail to apply for coverage at the appropriate time. Virtually all individuals who are 66 or older are covered by Medicare.

In addition, most individuals make a choice of supplemental private insurance coverage shortly before or in their 65th year because the price of such insurance rises sharply with age and coverage becomes more restrictive. The response variable is the number of visits to a physician in an office setting (denoted by ofp in what follows). Available covariates include: i) socio-economic variables: gender (1 for female, 0 for male), age (in years, divided by 10), marital status, educational level (number of years of education), income, ii) various measures of health condition: number of chronic conditions (cancer, arthritis, gallbladder problems $\cdots$) and a variable indicating self-perceived health level (poor, average, excellent) and iii) a binary variable indicating whether individual is covered by medicaid or not (medicaid is a US health insurance for individuals with limited income and resources, we code it as 1 if the individual is covered and 0 otherwise). Self-perceived health is recoded as two dummy variables denoted by "health1" (1 if health is perceived as poor, 0 otherwise) and "health2" (1 if health is perceived as excellent, 0 otherwise).

We fit the following four models : i) a ZIP regression model where $\lambda_i$ and $\pi_i$ are specified as in (2.3)-(2.4); ii) the ZIP model with cloglog link (denoted by ZIP-cloglog thereafter)where $\pi_i$ is as in (2.10) , iii) the ZIP model with with probit link (denoted by ZIP-probit), where $\pi_i$ is as in (2.12) and iv) the ZIP model with GEV link (denoted by ZIP-GEV),where $\pi_i$ is as in (2.7). Selection of regressors for inclusion in $\pi_i$ requires some care. Indeed, it was previously observed in various other zero-inflated models that including all available regressors in both count and zero-inflation probabilities can yield lack of identification of model parameters. See for example Diop et al. (2011) and Staub et al (2013), who suggest to solve this issue by letting at least one of the covariates included in the count model to be excluded from the zero-inflation model (or the converse). Such condition is not required in the ZIP model.

Using the Wald testing, we identify five significant predictors : age, gender, educational level, number of chronic conditions and medicaid status, that are included in $\pi_i$.

Results for the fours competing models (ZIP, ZIP-cloglog , ZIP-probit and ZIP-GEV ) are displayed in Table 8. We report estimate, standard error and significance level of Wald test for each parameter. For purpose of comparison, we also report AIC and BIC values for the four models. ZIP-GEV appears as the best model in terms of both AIC . A closer look at the results from the widely used logit link regression model in the healthcare utilization research and our GEV regression model reveals some difference in the estimation of the covariates effects. Gender, educational level and medicaid status are identified by ZIP-GEV as the most influencing factors for being a permanent non-user, with medicaid recipients being more likely to be permanent non-users. The four models identify the same subset of influent factors for healthcare utilization, with similar parameter estimates.

From Table 8, we observe that in the overall population, significant determinants of the decision to consult a non-physician when visiting in an office setting include health status, age, gender, educational level and medicaid status. Patients with poor health will favor office visits to a physician over office visits to a non-physician, which seems a natural observation. Women and people with higher education have higher probability to consult a non-physician, while medicaid recipients are more likely to visit physicians than non-physicians. The probability of visiting a non-physician when consulting in an office setting decreases with age. This may be due to several factors, such as decreasing mobility associated with ageing (aged patients will tend to limit their consultations to those considered as the most necessary, that is, to physician visits) and worsening of the health condition with ageing (patients whose health declines are likely to favor visits to a physician).

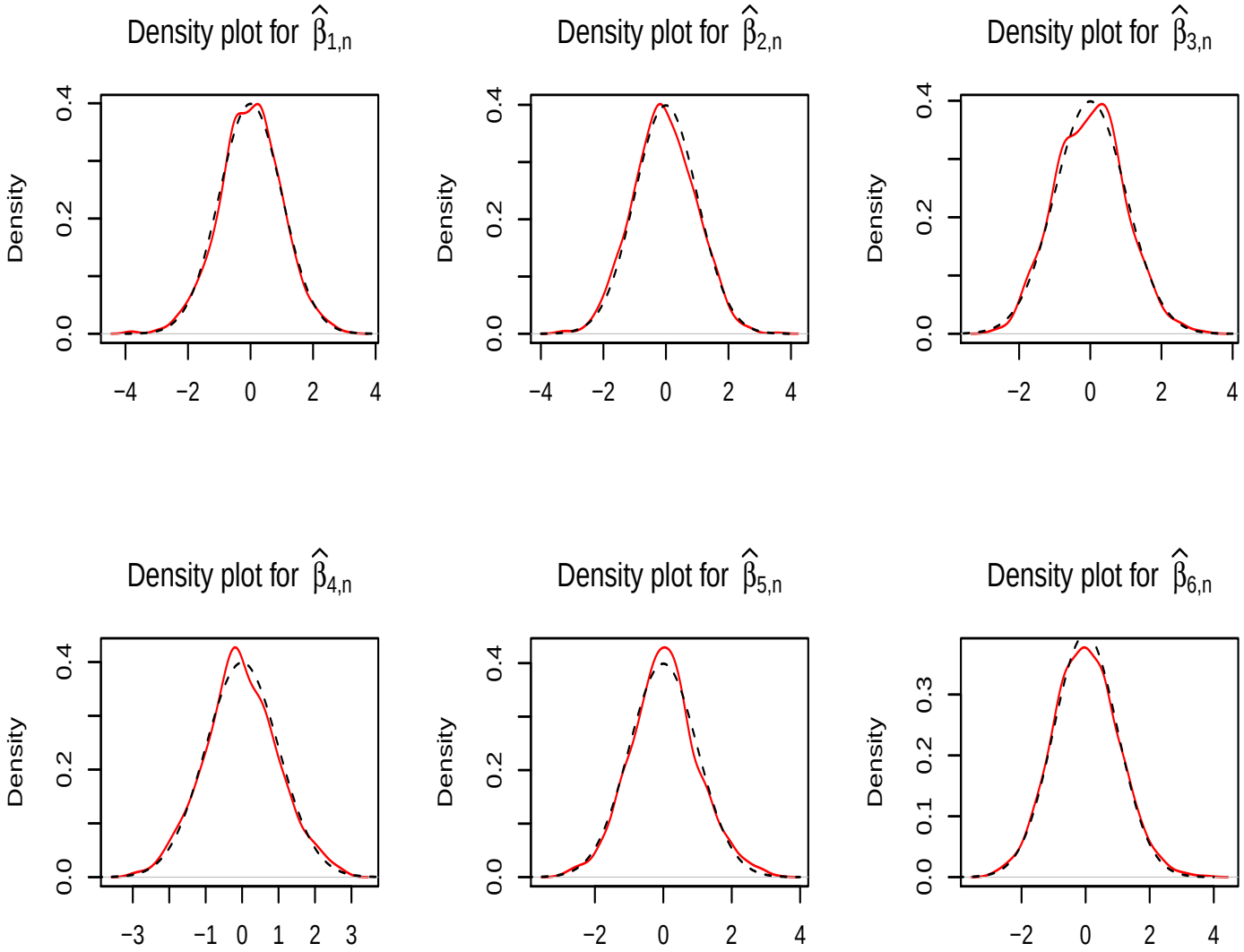


Figure 1: Density estimates of the $(\hat{\beta}_{j,n} - \beta_j)/\text{standard error}(\hat{\beta}_{j,n}), j = 1, \ldots, 6$ with $n = 500$ and 25% of zero-inflation. using the ZIP-GEV model

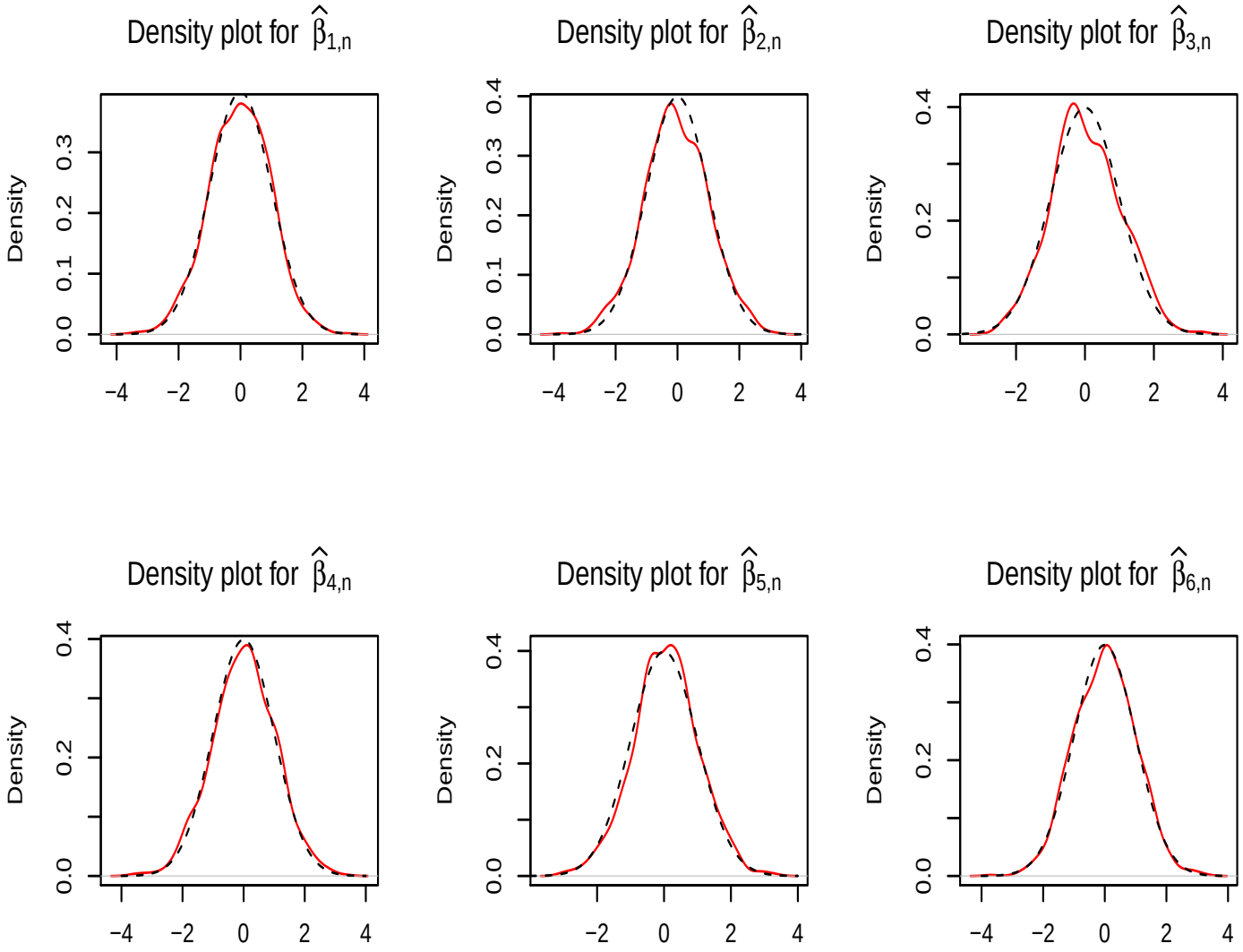


Figure 2: Density estimates of the $(\hat{\beta}_{j,n} - \beta_j)/\text{standard error}(\hat{\beta}_{j,n}), j = 1, \ldots, 6$ with $n = 500$ and 20% of zero-inflation. using the ZIP-probit model

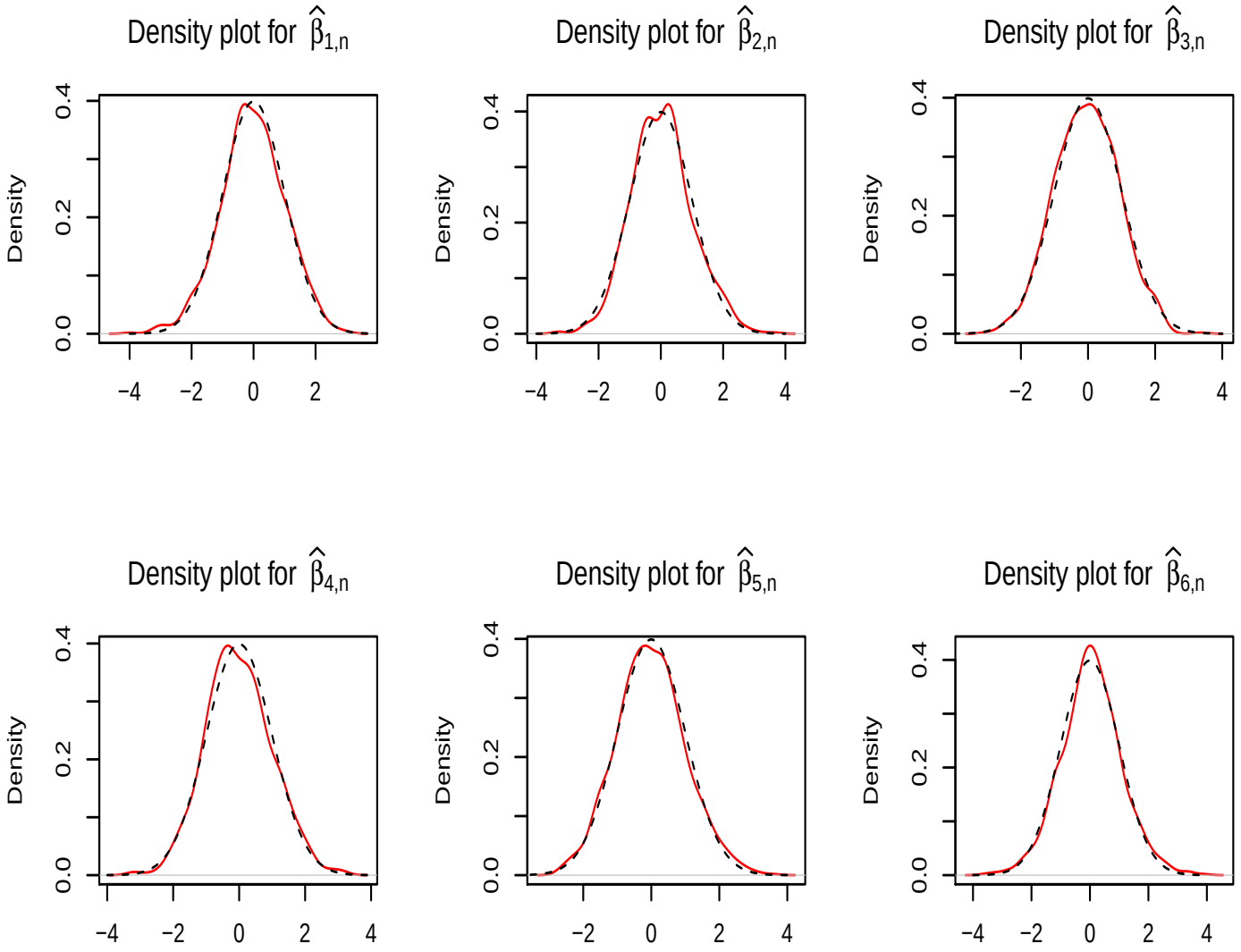


Figure 3: Density estimates of the $(\hat{\beta}_{j,n} - \beta_j)$/standard error$(\hat{\gamma}_{k,n}), k = 1, \ldots, 6$ with $n = 500$ and 30% of zero-inflation. using the ZIP-cloglog model

| parameter | | ZIP | | | ZIP-cloglog | | | ZIP-probit | | | ZIP-GEV | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | est | s.e | signif. | est | s.e | signif. | est | s.e | signif. | est | s.e | signif. |
| $\beta$ | intercept | 1.14 | 0.082 | *** | 0.0953 | 0.0801 | | 0.9717 | 0.082 | *** | 1.6778 | 0.0827 | *** |
| | health1 | 0.2199 | 0.0172 | *** | 0.1327 | 0.0177 | *** | 0.1507 | 0.0178 | *** | 0.2037 | 0.0175 | *** |
| | health2 | -0.2883 | 0.0321 | *** | -0.1600 | 0.0294 | *** | -0.3634 | 0.0324 | *** | -0.2196 | 0.0300 | *** |
| | chronic | 0.1053 | 0.0047 | *** | 0.0778 | 0.0047 | *** | 0.098 | 0.0047 | *** | 0.0981 | 0.0046 | *** |
| | age | 0.0759 | 0.0104 | *** | 0.2134 | 0.0101 | *** | 0.1069 | 0.0104 | *** | 0.0145 | 0.0106 | |
| | gender | 0.1544 | 0.0147 | *** | 0.2080 | 0.0146 | *** | 0.0554 | 0.0146 | *** | 0.0379 | 0.0144 | ** |
| | marital status | -0.0157 | 0.0145 | | 0.0822 | 0.0144 | *** | -0.0017 | 0.0146 | | -0.0008 | 0.0019 | |
| | medicaid | -0.2314 | 0.0125 | *** | 0.0412 | 0.0087 | *** | -1.2315 | 0.0801 | *** | -1.8564 | 0.0450 | *** |
| $\gamma$ | intercept | -1.1479 | 0.419 | ** | -0.2993 | 0.1503 | * | 0.1404 | 0.3913 | *** | 0.2124 | 0.5279 | *** |
| | health1 | -0.4416 | 0.3049 | | 0.6684 | 0.0974 | *** | -1.0049 | 0.4719 | * | -0.8024 | 0.4908 | |
| | gender | -0.0465 | 0.2945 | | 0.4303 | 0.1115 | *** | -1.3877 | 0.4375 | ** | -1.9849 | 0.4939 | *** |
| | marital status | 2.3482 | 0.4095 | *** | -0.1539 | 0.1044 | | 1.3452 | 0.3823 | *** | 1.8333 | 0.4916 | *** |
| | education | -0.531 | 0.0365 | *** | -0.1642 | 0.0087 | *** | -1.1599 | 0.2616 | *** | -1.5664 | 0.5250 | ** |
| | AIC | | -49683.08 | | | -49097.52 | | | -49709.5 | | | **-49860.05** | |
| | BIC | | -9015.084 | | | -8429.517 | | | -9041.499 | | | **-9192.049** | |

Table 8: : Health-care data analysis: estimates,standard errors and significance codes: *** significant at the 0.1% level, ** significant at the 1% level, * significant at the 5% level, . significant at the 10% level..

## 5. Concluding Remarks

In this paper, we study the properties of MLE in ZIP regression models when the susceptibility probability function is modeled with different links functions. Our simulations suggest that the MLE works well and that reliable statistical inferences about the parameters of interest in the different models can be based on the normal approximation of the MLE distribution. Maximum likelihood estimation is shown to perform well in this model, under a range of scenarios. Moreover, in our analysis of health-care utilization, the proposed model provides plausible explanations and interpretations and gives useful insight of the decision of using or not available healthcare services. Several issues now require more attention, such as estimation in the bivariate ZIP-GEV regression in various forms. Investigating the estimation of a flexible ZIP regression model that combines a generalized extreme value link function with a Gaussian process is also desirable. All these issues will be tackled in future works.

## References


Agresti, A., 2002. Categorical Data Analysis. Wiley, New York.

Ali, E., Diop, A. and Dupuy, J.-F., 2020. A constrained marginal zero-inflated binomial regression model. Communications in Statistics-Theory and Methods, doi: 10.1080/03610926.2020.1861296

Aranda-Ordaz , F. J., 1981. On two families of transformations to additivity for binary response data. Biometrika 68 357-363.

Abid, R.,Kokonendji, C.C., and Masmoudi, A., 2020. Geometric Tweedie regression models for continuous and semicontinuous data with variation phenomenon. AStA. Adv. Statist. Anal. 104,pp. 33-58.

Bonat,W.H., Jorgensen, B., Kokonendji, C.C., Hinde, J. , and Demetrio C.G.B., 2018. Extended Poisson-Tweedie: properties and regression models for count data, Stat. Model. 18 , pp. 24-49

Calabrese, R. and Osmetti, S. A., 2013. Modelling Small and Medium Enterprise Loan Defaults as Rare Events: The Generalized Extreme Value Regression Model. Journal of Applied Statistics, 40(6), 1172-1188.

Chen, M.-H., Dey, D. K. and Shao, Q.-M. 1999 A new skewed link model for dichotomous quantal response data. J. Amer. Statist. Assoc. 94 1172-1186.

Coles S. G. 2004. An Introduction to Statitical Modelling of Extreme Values. Springer-Verlag, London.

Czado, C. and Santner, T. J. 1992. The effect of link misspecification on binary regression inference. J. Statist. Plann. Inference 33 213-231.

Czado, C., Erhardt, V., Min, A., Wagner, S. 2007. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. Statistical Modelling 7(2), 125-153.

Czado, C., Min, A., 2005. Consistency and asymptotic normality of the maximum likelihood estimator in a zero-inflated generalized Poisson regression. Collaborative Research Center 386, Discussion Paper 423 , Ludwig-Maximilians-Universität, München.

Chen, M.-H. and Shao, Q.-M., 2000. Propriety of posterior distribution for dichotomous quantal response models with general link functions. Proc. Amer. Math. Soc. 129 293-302.

Coles, S. G., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, New York.

Deb, P., Trivedi, P. K., 1997. Demand for medical care by the elderly: a finite mixture approach. Journal of Applied Econometrics 12(3), 313-336.

Deng, D., Zhang, Y., 2015. Score tests for both extra zeros and extra ones in binomial mixed regression models. Communications in Statistics - Theory and Methods 44, 2881-2897.

Diallo, A. O., Diop, A., Dupuy, J.-F., 2017. Asymptotic properties of the maximum likelihood estimator in zero-inflated binomial regression. Communications in Statistics - Theory and Methods 46(20), 9930-9948.

Dietz, E., Böhning, D., 2000. On estimation of the Poisson parameter in zero-modifed Poisson models. Computational Statistics & Data Analysis 34(4), 441-459.

Diop, A., Diop, A., Dupuy, J.-F., 2011. Maximum likelihood estimation in the logistic regression model with a cure fraction. Electronic Journal of Statistics 5, 460-483.

Diop, A., Diop, A., Dupuy, J.-F., 2016. Simulation-based inference in a zero-inflated Bernoulli regression model. Communications in Statistics - Simulation and Computation 45(10), 3597-3614.

Diallo, A. O., Diop, A., Dupuy, J.-F., 2018. Analysis of multinomial counts with joint zero-inflation, with an application to health economics. Journal of Statistical Planning and Inference 194, 85-105.

Diallo, A. O., Diop, A., Dupuy, J.-F., 2019. Estimation in zero-inflated binomial regression with missing covariates. Statistics 53(4), 839-865.

Dupuy, J.-F., 2018. Statistical Methods for Overdispersed Count Data. ISTE Press - Elsevier.

Eicker, F., 1966. A multivariate central limit theorem for random linear vector forms. The Annals of Mathematical Statistics 37(6), 1825-1828.

Feng, J., Zhu, Z., 2011. Semiparametric analysis of longitudinal zero-inflated count data. Journal of Multivariate Analysis 102, 61-72.

Foutz, R. V., 1977. On the unique consistent solution to the likelihood equations. Journal of the American Statistical Association 72, 147-148.

Guerrero,V.M.and Johnson, R. A. 1982. Use of the Box-Cox transformation with binary response models. Biometrika 69 309-314.

Hall, D. B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56(4), 1030-1039.

He, X., Xue, H., Shi, N.-Z.,2010. Sieve maximum likelihood estimation for doubly semiparametric zero-inflated Poisson models. Journal of Multivariate Analysis 101, 2026-2038.

Henningsen, A., Toomet, O., 2011. maxLik: A package for maximum likelihood estimation in R. Computational Statistics 26(3), 443-458.

Johnson,N.L., Kemp, A.W. and Kotz S., 2005 Univariate Discrete Distributions, 3rd ed., Wiley, New York,

Jiang, J., 2010. Large Sample Techniques for Statistics. Springer, New York.

Jones, M. C. 2004. Reply to Comments on "Families of distributions arising from distributions "of order statistics. TEST 13 1-43.

Kokonendji, C.C. 2014. Over- and underdispersion models, in The Wiley Encyclopedia of Clinical Trials- Methods and Applications of Statistics in Clinical Trials, N. Balakrishnan, ed., Vol. 2, Chapter 30, Wiley, New York, pp. 506-526

Lam, K. F., Xue, H., Cheung, Y. B., 2006. Semiparametric analysis of zero-inflated count data. Biometrics 62(4), 996-1003.

Lim, H. K., Li, W. K., Yu, P. L. H.,2006 . Zero-inflated Poisson regression mixture model. Computational Statistics & Data Analysis 71, 151-158.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1-14.

Long, D. L., Preisser, J. S., Herring, A. H., Golin, C. E., 2014. A marginalized zero-inflated Poisson regression model with overall exposure effects. Statistics in medicine 33(29), 5151-5165.

Martin, J., Hall, D. B., 2017. Marginal zero-inflated regression models for count data. Journal of Applied Statistics 44(10), 1807-1826.

Moghimbeigi, A., Eshraghian, M. R., Mohammad, K., McArdle, B., 2008. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. Journal of Applied Statistics 35(9), 1193-1202.

McCullagh, P., Nelder, J. A., 1989. Generalized linear models (Second edition). Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Monod, A., 2014. Random effects modeling and the zero-inflated Poisson distribution. Communications in Statistics. Theory and Methods 43(4), 664-680.

Min, Y., Agresti,A., 2005. Random effect models for repeated measures of zero-inflated count data. Statistical Modelling 5(1), 1-19.

Nguyen, V. T., Dupuy, J.-F.,2019. Asymptotic results in censored zero-inflated Poisson regression. Springer Series in Statistics, Springer.

Preisser, J. S., Stamm, J. W., Long, D. L., Kincade, M. E., 2012. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries research 46(4), 413-423.

R Core Team, 2018. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

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

Stukel,T. A., 1988. Generalized logistic models. J. Amer. Statist. Assoc. 83 426-431.

. Consistent estimation of zero-inflated count models. Health Economics, 22(6) :673-686, 2013.

Smith, R. L., 2003. Statistics of extremes, with applications in environment, insurance and finance. In Extreme Values in Finance, Telecommunications and the Environment (B. Finkenstadt and H. Rootzen, eds.) 1-78. Chapman and Hall/CRC Press, London.

Todem, D., Kim, K., Hsu, W. W., 2016. Marginal mean models for zero-inflated count data. Biometrics 72(3), 986-994.

Wang, X. and Dey, D. K., 2010. Generalized extreme value regression for binary response data:An application to B2B electronic payments system adoption. Ann. Appl. Stat. 4 2000-2023.

Wu, Y., Chen, M.-H. and Dey, D., 2002. On the relationship between links for binary response data. J. Stat. Stud. Special Volume in Honour of Professor Mir Masoom Ali's 65th Birthday 159-172.