Computational Statistics & Data Analysis 34 (2000) 441–459

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

On estimation of the Poisson parameter in zero-modified Poisson models

Ekkehart Dietz*, Dankmar Böhning

Department of Epidemiology, Free University Berlin, Haus 562, Fabeckstr. 60-62, 14195 Berlin, Germany

Received 1 February 1999; received in revised form 1 September 1999; accepted 26 November 1999

Abstract

For count data, typically, a Poisson model is assumed. However, it has been observed in various applications that this model does not fit, because of too few or too many zeros in the data. For the latter case the zero-inflated Poisson (regression) model has been proposed. In situations where, for certain reasons, no zeros at all can be observed, the zero-truncated Poisson (regression) model is appropriate. This paper provides an EM algorithm for ML estimation of the latter model by standard software for Poisson regression. It is shown, that this algorithm can be used also to estimate the Poisson parameters of zero-inflated, zero-deflated, and standard Poisson models, when the zero observations are ignored. If nothing is known about the kind of zero modification, the respective estimates are fully efficient. Situations are described, in which the loss of efficiency is small although knowledge of the kind of zero modification is available. In a second step, the respective estimates of the Poisson parameter can be used to analyze the kind of zero modification and to estimate the zero modification parameter. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Zero-modified Poisson model; Zero-truncated Poisson model; Zero-inflated Poisson model; Zero-deflated Poisson model

1. Introduction

The analysis of count data occupies a prominent place in applied statistics in many areas. Sometimes, such data are not Poisson distributed, although a Poisson process as source of the data seems to be plausible. In such cases this discrepancy very often

* Corresponding author.

PII: S 0167-9473(99)00111-5

concerns mainly the zero-count class. That is, too many or too few zeros occur. This may happen in the following situations:

Situation 1: Not all members of the study population considered are affected by the Poisson process, so that *zero inflation* occurs because of zero response of the unaffected members.

Situation 2: Certain unavoidable problems in the sampling process lead to an increased or decreased chance of the zero-count members of the population coming into the sample. This leads to *zero inflation* or *zero deflation*, respectively.

Situation 3: As an extreme case of Situation 2, there is no chance at all of getting a zero observation into the sample. This is called *zero truncation*. The respective distribution of the sample data is called *positive Poisson distribution*.

Situation 4: As a combination of Situations 1 and 3, we have a subpopulation with a data generating process leading to a positive Poisson distribution, whereas the complementary population, which is not affected by this process, provides zero-count observations.

The problem of zero modification in count data has a long history in the statistical literature. In McKendrick (1926), David and Johnson (1952), and Gurmu (1991), the ML estimation of zero-truncated Poisson distributions is considered. In the first paper, a simple iterative estimating procedure is given, where in each iterate, the expected sample size of a respective non-zero-truncated sample is computed. This quantity may be of special interest in some applications.

In Cohen (1960) and Umbach (1981), the ML estimation of a zero-truncated Poisson model is used as estimation of the Poisson parameter of more general zero-modified Poisson distributions. In this stage, they simply ignore the zero observations in the sample. The latter ones are only used to estimate the additional model parameter in a second stage. Lambert (1992) introduced a zero-inflated Poisson regression model and Xie and Aickin (1997) applied a zero-truncated Poisson regression model.

In this article, a general zero-modified Poisson regression model is considered. In Section 2, generalized Poisson model families are defined, which are appropriate in the situations described above. In Section 3, it is shown that its Poisson parameters can be estimated without using the observed zeros and knowledge of the kind of violation of Poisson assumption. In the subsequent sections, the loss of efficiency of maximum likelihood estimation is studied, when doing so.

Example data. For illustration, we will consider the data from a dental epidemiological study published by Mendonca (1995). Fig. 1 displays the histogram of the DMFT index of 797 children being all 7 years old and living in an urban area of Belo Horizonte (Brazil). The DMFT index is an important and well-known overall measurement of the dental status of a person. It is a count number standing for the number of Decayed, Missing, and Filled Teeth. The original aim of this study was to estimate the effect of four prevention measures applied in five of the six schools from which the children were recruited. No explicit prevention measure was applied to children in the sixth school. The children in this school serve as control group. Interested readers can look up Mendonca (1995).

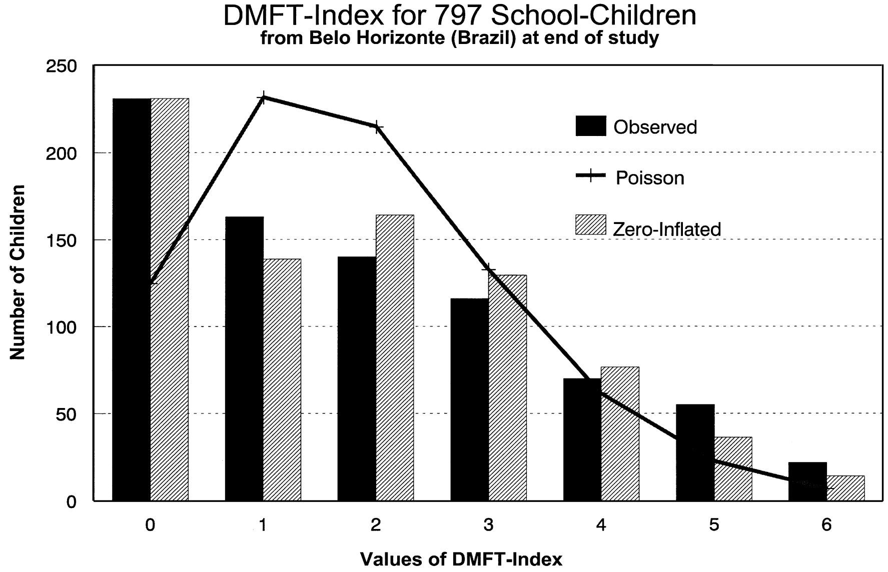

Fig. 1. Values of DMFT-index.

As seen in Fig. 1, there is an obvious zero inflation in the data if compared with a standard Poisson distribution. The ML estimation of a zero-inflated Poisson distribution provides a clearly improved fit of the data. Based on the model family considered in this paper, we will study the influence of the prevention measures on both the Poisson parameter and the zero inflation.

2. Zero-modified Poisson models

2.1. Zero-modified Poisson distribution

When no covariates have to be considered, for situations described above, we can use the following model family: Let y be a response variable, where $y \in \{0, 1, 2, \ldots\}$. Its distribution is defined by its probability density function as

$$f_{ZMP}(y, \mu, p) = (1 - p) f_{\mathrm{Po}}(y, 0) + p f_{\mathrm{Po}}(y, \mu), \tag{1}$$

where $f_{\mathrm{Po}}(y, \mu)$ denotes the pdf of the Poisson distribution, μ its mean parameter, and $P(y, 0)$ its degenerate variant with $\mu = 0$, which gives all mass to zero. $f_{\mathrm{Po}}(y, 0) = 1$ if $y = 0$ and $f_{\mathrm{Po}}(y, 0) = 0$ otherwise. The parameter μ is assumed to be positive. For the additional parameter p, it is presupposed, that

$$0 \le p \le \frac{\mathrm{e}^{\mu}}{\mathrm{e}^{\mu} - 1}$$

holds. This condition makes sure that the right-hand side of (1) is a pdf. This model family is denoted as zero-modified Poisson distribution. It is represented by $ZMP(\mu, p)$. Its additional parameter p is called *zero-modification parameter*.

Different values of p lead to different modifications of the Poisson model, where the proportion of additional or of missing zeros, respectively, is

$$1 - p + p\mathrm{e}^{-\mu} - \mathrm{e}^{-\mu} = (1 - p)(1 - \mathrm{e}^{-\mu}). \tag{2}$$

$p = 0$: In this case (1) degenerates to a one-point distribution giving all mass on zero.

$0 < p < 1$: This yields a zero-inflated Poisson distribution, which is a Poisson distribution with a proportion of additional zeros. Such a distribution may be appropriate in Situations 1, 2, and 4 as described in the previous section.

$p = 1$: This yields the usual Poisson distribution without any modification. Eq. (2) becomes zero in this case.

$1 < p < \mathrm{e}^{\mu}/(\mathrm{e}^{\mu} - 1)$: In this case, (2) becomes negative. That is, less zeros occur, than expected under the Poisson distribution. Such models are denoted as zero-deflated Poisson distribution. Models like this may fit in Situations 2 and 4.

$p = \mathrm{e}^{\mu}/(\mathrm{e}^{\mu} - 1)$: In this case, (1) becomes the zero-truncated Poisson distribution, where the parameter p cancels out and no longer appears as a model parameter. The pdf of the ZTP distribution $ZTP(\mu)$ is

$$f_{ZTP}(y, \mu) = \mathrm{e}^{-\mu}\mu^{y}/[y!(1 - \mathrm{e}^{-\mu})], \quad y = 1, 2, 3, \ldots . \tag{3}$$

It is appropriate in Situation 3 of Section 1.

2.2. Zero-modified Poisson regression

Model (1) can be generalized by a ZMP regression model, which allows both the Poisson parameter μ and the weight parameter p to vary. Thereby, at least, the variation of μ is assumed to depend on a vector of covariables x. A quite general ZMP regression model can be defined as

$$y \mid x \sim ZMP(\mu(\beta_1^{\mathrm{T}}x), p), \tag{4}$$

where p is an unobserved variable fulfilling the condition

$$0 < p < \frac{\mathrm{e}^{\mu(\beta_1^{\mathrm{T}}x)}}{\mathrm{e}^{\mu(\beta_1^{\mathrm{T}}x)} - 1},$$

β_1 is a parameter vector, $\mu(\beta_1^{\mathrm{T}}x) = g^{-1}(\beta_1^{\mathrm{T}}x)$, and $g(.)$ is a suitable link function. In this paper, we assume the log link as usual in standard Poisson regression. Applying the parameter restrictions given above, one could obtain the definitions of a Poisson regression model, a ZIP regression model, and a ZDP regression model, respectively. To get sensible predictions also of the unobserved variable p of (4) one has to assume further parameter restrictions. This can be done by another linear predictor and a suitable link function

$$p(\beta_2^{\mathrm{T}}x) = g_p^{-1}(\beta_2^{\mathrm{T}}x).$$

Depending on the kind of data generating processes considered, different link functions may be appropriate. For the ZIP regression, the respective parameter restriction is fulfilled, when the logit link is chosen (Lambert, 1992), that is

$$g_p(p) = \log(p/(1 - p)).$$

A ZDP regression model is obtained, when choosing

$$g_p(p) = \log\left((e^{\mu} - 1) \Big/ \left(\frac{p}{p-1} - e^{\mu}\right)\right),$$

where $\mu = \mu(\beta_1^{\mathrm{T}} x)$. That is, the Poisson parameter serves as an additional parameter of the link function. We now have a situation in which a link function parameter depends on covariables. It is easily possible to implement such a link in GLIM software. The link function above guarantees that

$$1 < p(\beta_2^{\mathrm{T}} x) < \frac{e^{\mu}}{e^{\mu} - 1}.$$

In general, however, model (4) should allow different kinds of zero modification for different values of the covariables within the same model. As a reasonable choice of the link function for the weight parameter in situation 4, we consider here

$$g_p(p) = g_{\mu}(p) = \log\left(\frac{p}{(e^{\mu}/(e^{\mu} - 1) - p)}\right)$$

which leads to

$$p(\beta_2^{\mathrm{T}} x) = g_{\mu}^{-1}(\beta_2^{\mathrm{T}} x) = \frac{e^{\beta_2^{\mathrm{T}} x}}{1 + e^{\beta_2^{\mathrm{T}} x}} \frac{e^{\mu}}{e^{\mu} - 1}. \tag{5}$$

This link function guarantees that

$$0 < p(\beta_2^{\mathrm{T}} x) < \frac{e^{\mu}}{e^{\mu} - 1},$$

so that (4) is a regression model. It practically excludes the case of zero-truncated Poisson regression, which is the degenerate model, if the linear predictor $\beta_2^{\mathrm{T}} x$ goes to infinity for each value of x. In this case, the first factor of the right-hand side of (5) is equal to one. Replacing p in (4) by

$$p(\beta_2^{\mathrm{T}} x) = g_{\mu}^{-1}(\beta_2^{\mathrm{T}} x) = \frac{e^{\mu}}{e^{\mu} - 1}$$

leads to

$$y \,|\, x \sim f_{ZTP}(y, \mu(\beta_1^{\mathrm{T}} x)). \tag{6}$$

The inverse link function (5) models the probability of a zero response as the inverse logit function of the same predictor. This becomes clear, if another representation of the ZMP regression model is considered. We display the right-hand side of (4) as

$$\begin{aligned}&[(1 - p(\beta_2^{\mathrm{T}} x)) + p(\beta_2^{\mathrm{T}} x) * f_{\mathrm{Po}}(0, \mu(\beta_1^{\mathrm{T}} x)] * f_{\mathrm{Po}}(y, 0)\\ &\quad + [p(\beta_2^{\mathrm{T}} x) * f_{\mathrm{Po}}(y, \mu(\beta_1^{\mathrm{T}} x))] * (1 - f_{\mathrm{Po}}(y, 0)),\end{aligned}$$

where the first term gives the probability of zero and the second term gives the probability of a positive outcome y. Thus, we have a reparametrization of $f_{ZMP}(y, \mu, p)$ by

$$f'_{ZMP}(y, \mu, p') = (1 - p'(\beta_2^{\mathrm{T}} x)) * f_{\mathrm{Po}}(y, 0) + p'(\beta_2^{\mathrm{T}} x) * f_{ZTP}(y, \mu(\beta_1^{\mathrm{T}} x)), \tag{7}$$

where

$$p'(.) = p(.)(1 - \mathrm{e}^{-\mu}). \tag{8}$$

$1 - p'(\beta_2^{\mathrm{T}} x)$ is just the conditional probability of zero in the ZMP model, where the respective parameter restriction can now be written as

$$0 < p'(\beta_2^{\mathrm{T}} x) < 1.$$

Replacing $p(.)$ by (5) leads to $p'(\beta_2^{\mathrm{T}} x) = \mathrm{e}^{\beta_2^{\mathrm{T}} x}/(1 + \mathrm{e}^{\beta_2^{\mathrm{T}} x})$.

3. A two-step estimation of a ZMP model

The following two-step procedure is motivated by the fact that zero-truncated ZMP models are always ZTP models having the same Poisson parameter or linear predictor, respectively. This is easy to see when representation (7) of the ZMP regression model is used. The conditional probability of a positive response is

$$P(y > 0 \,|\, x) = p'(\beta_2^{\mathrm{T}} x).$$

Excluding the zero from the domain of (7) and dividing its right-hand side by the quantity above leads immediately to

$$y \,|\, x \sim f_{ZTP}(y, \mu(\beta_1^{\mathrm{T}} x)).$$

As a consequence, the Poisson parameter or its predictor, respectively, can be estimated consistently from the "positive" data only.

So, in the first step of our procedure, we will ignore the data records with a zero outcome and we will compute the ML estimate of the Poisson parameter based on a zero-truncated Poisson model. This can be performed easily by standard software (see Section 4).

In the second step, the zero-modification parameter or its linear predictor is estimated from the whole data set, where the Poisson parameters μ_i are assumed to be equal to the estimate obtained in the first step. This can be done by a conditional ML estimation procedure as described in the next section. Let p_i and μ_i be abbreviations of $p(\beta_2^{\mathrm{T}} x_i)$ and $\mu(\beta_1^{\mathrm{T}} x_i)$, respectively. If $p_i = p \;\forall i$ is assumed, p can also be estimated be solving the equation

$$\sum_{i=1}^{n} p_i' = \sum_{i=1}^{n} p(1 - \mathrm{e}^{-\hat{\mu}_i}) = n - n_0,$$

where $\hat{\mu}_i$ is an abbreviation of $\mu(\hat{\beta}_1^{\mathrm{T}} x_i)$, n denotes the sample size, and n_0 denotes the number of sample elements having a zero as the count data. The left-hand side of this equation is just the expected number of positive counts in the sample given the μ_i. This leads to the estimate

$$\hat{p} = \frac{n - n_0}{\sum_{i=1}^{n} (1 - \mathrm{e}^{-\mu_i})}. \tag{9}$$

In the most simplest case, if μ is constant and does not depend on covariables, the equation above simplifies to

$$\hat{p} = \frac{n - n_0}{n(1 - \mathrm{e}^{-\mu})}. \tag{10}$$

What remains is the question: What is the loss of efficiency one has to take into the bargain, when ignoring the zeros in the data in the first step of this procedure. In this article, we try to give an answer with respect to the maximum likelihood estimation of the Poisson parameters or its log linear predictors, respectively.

4. Maximum likelihood estimation

4.1. ZTP distribution

Let n_j denote the number of responses j ($j = 0, 1, 2, \ldots$) in sample data. The log likelihood function of the ZTP distribution for the positive responses of the same sample is

$$\begin{aligned} l(\mu) &= \sum_{j=1}^{\infty} n_j \log[f_{ZTP}(\mu, j)] \\ &= \sum_{j=1}^{\infty} n_j \log[f_{\mathrm{Po}}(\mu, j)/(1 - \mathrm{e}^{-\mu})] \end{aligned} \tag{11}$$

leading to the likelihood equation

$$-(n - n_0) + \frac{\bar{y}_p}{\mu}(n - n_0) - \frac{\mathrm{e}^{-\mu}}{1 - \mathrm{e}^{-\mu}}(n - n_0) = 0,$$

where n denotes the sample size and $\bar{y}_p$ the mean of the positive y_i. The equation above can be written as

$$\mu = \bar{y}_p(1 - \mathrm{e}^{-\mu}) =: G(\mu). \tag{12}$$

Because

$$\frac{\partial}{\partial \mu} G(\mu) = \bar{y}_p \mathrm{e}^{-\mu} > 0,$$

$\mu_{l+1} = G(\mu_l)$ converges for any initial value μ_0 to the ML estimate satisfying the fixed point equation $\mu = G(\mu)$. The only assumption which has to be made is that at least one sample element is positive.

The fixed point equation (12) can be written as

$$\mu = \left[\sum_{i=1}^{n} y_i\right] * \left[\frac{1 - \mathrm{e}^{-\mu}}{n_+}\right], \tag{13}$$

where n_+ denotes the number of positive observations. Notice that only positive y_i's provide a contribution to the first term of the right-hand side of this equation. The second term is just the inverse of the expected sample size, if only μ and the number of positive sample elements n_+ is known. Notice that the conditional number of zero observations given μ and n_+ is distributed as $NB(n_+, 1-\mathrm{e}^{-\mu})$, where $NB(.,.)$ denotes the negative binomial distribution. Its expectation is

$$E(n_0 \mid \mu, n_+) = \frac{n_+\mathrm{e}^{-\mu}}{1-\mathrm{e}^{-\mu}}.$$

Thus, the conditional expectation of n is

$$n_+ + \frac{n_+\mathrm{e}^{-\mu}}{1-\mathrm{e}^{-\mu}} = \frac{n_+}{1-\mathrm{e}^{-\mu}}$$

and (13) can be written as

$$\mu = \left[\sum_{i=1}^{n} y_i\right] \bigg/ (n_+ + E(n_0 \mid \mu, n_+)).$$

From this, it can be seen, that the iterative application of (12) can also be considered as an EM algorithm, where in the E-step the expected number of zeros in the respective non-truncated Poisson distribution is computed. The M-step is the usual ML estimator of this Poisson distribution, applied to the positive observations augmented with the expected number of zeros.

4.2. ZMP distribution vs. ZTP distribution

4.2.1. ML estimation of ZMP distributions

The log likelihood function of a ZMP distribution having constant $p_i = p$ is

$$l(p,\mu) = n_0 \log[(1-p) + p\mathrm{e}^{-\mu}] + \sum_{j=1}^{\infty} n_j \log[p f_{\mathrm{Po}}(j,\mu)].$$

From the *score vector*

$$\left(n_0\frac{\mathrm{e}^{-\mu}-1}{1-p+p\mathrm{e}^{-\mu}} + \frac{n-n_0}{p}, n_0\frac{-p\mathrm{e}^{-\mu}}{1-p+p\mathrm{e}^{-\mu}} - (n-n_0)\left(1-\frac{\bar{y}_p}{\mu}\right)\right)^{\mathrm{T}},$$

one obtains the score equations

$$\mu = \frac{\bar{y}_p}{p}\frac{n-n_0}{n} \tag{14}$$

and

$$p = \frac{1}{1-\mathrm{e}^{-\mu}}\frac{n-n_0}{n} \tag{15}$$

which can be written in one equation, which is identical to the fixed point equation (12). That is, ML-estimation of the ZMP model can be obtained by the ML estimation of the ZTP model based on the positive response data.

After estimating the Poisson parameter in this way, in a second step, Eq. (15) can be used also to get the ML estimation of the weight parameter p of the ZMP model. As additional information from the data, only the number of zero responses is needed in this step.

4.2.2. Relative efficiency

In the previous section, the following result has been proved:

Result 1. *The maximum likelihood estimate of the Poisson parameter μ of the ZMP distribution* (1) *can be obtained without loss of efficiency by the maximum likelihood estimate of the parameter of the ZTP distribution for the positive sample elements. Using this ML estimate, a fully efficient estimation of the parameter p can be obtained from the proportion of zero responses in the data.*

A loss of efficiency of the estimation based on the ZTP model can occur only if more is known about the kind of zero modification.

No zero modification: Let us assume, that the Poisson distribution is the true model, that is, the true value of p is 1. The asymptotic variance of the ML estimate of the Poisson parameter μ is

$$\sigma^2_{\hat{\mu}_{\mathrm{Po}}} = \frac{1}{n}\mu.$$

The asymptotic variance of the ML estimate of the Poisson parameter in the ZMP model as well as of the respective zero-truncated model are obtained as

$$\sigma^2_{\hat{\mu}_{ZTP}} = \sigma^2_{\hat{\mu}_{ZMP}} = \frac{1}{n-n_0}\frac{\mu(1-\mathrm{e}^{-\mu})^2}{1-(1+\mu)\mathrm{e}^{-\mu}}.$$

(see e.g. Xie and Aickin, 1997). Thus, one obtains the asymptotic relative efficiency as

$$ARE = \frac{\sigma^2_{\hat{\mu}_{\mathrm{Po}}}}{\sigma^2_{\hat{\mu}_{ZTP}}} = \frac{n-n_0}{n}\frac{1-(1+\mu)\mathrm{e}^{-\mu}}{(1-\mathrm{e}^{-\mu})^2}. \qquad (16)$$

Note that for a Poisson distribution

$$\frac{n-n_0}{n} \to 1-\mathrm{e}^{-\mu} \quad \text{as } n \to \infty.$$

Thus for n large enough, we have

$$ARE = \frac{1-(1+\mu)\mathrm{e}^{-\mu}}{1-\mathrm{e}^{-\mu}}.$$

The dotted line of Fig. 2 shows this asymptotic efficiency as function of the true Poisson parameter. The single points were obtained from 1000 simulated Poisson samples of size 100 for each of the parameter values $\mu = 0.5, 1, 2, 3, 4, 5, 6$.

As can be seen from this figure, the ZTP estimator is practically fully efficient, if the true parameter is larger than 8.

Now, let us assume a two-stage sampling procedure, where only units with positive outcome variable are taken into the sample. In some applications, such a procedure does not essentially increase the effort to obtain, say, a sample of size n. In such situations, it is sensible to consider another kind of relative efficiency, which compares the variance of the parameter estimate of a ZMP model obtained by usual sampling with the variance of the parameter estimate based on the ZTP model and a two stage

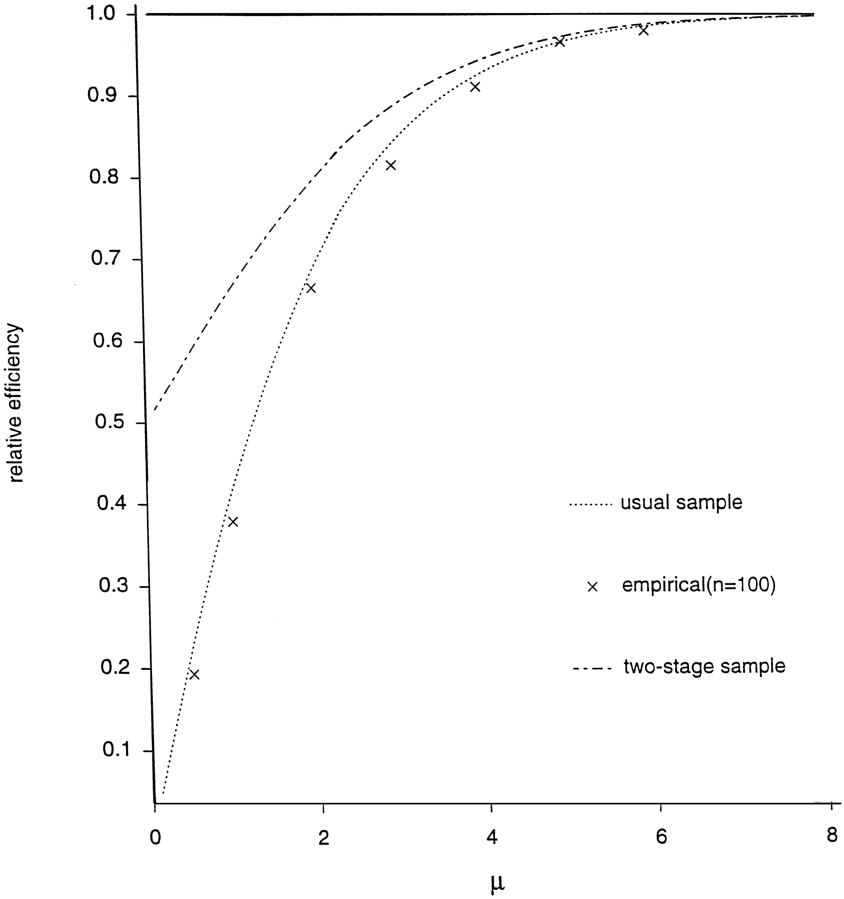

Fig. 2. Relative efficiency of the ZTP-estimation if a poisson distribution is true.

sample of the same size. For the Poisson model one obtains

$$ARE^{+} = \frac{\sigma^2_{\hat{\mu}_{\mathrm{Po}}}}{\sigma^{2+}_{\hat{\mu}_{ZTP}}} = \frac{1-(1+\mu)\mathrm{e}^{-\mu}}{(1-\mathrm{e}^{-\mu})^2}, \tag{17}$$

where $\sigma^{2+}_{\hat{\mu}_{\mathrm{ZTP}}}$ denotes the respective asymptotic variance bases on the two-stage sample of size n. The dashed line of Fig. 2 shows this relative efficiency as a function of the true Poisson parameter μ. As one can see, it is always larger than 0.5 and practically 1.0 for $\mu > 8$.

Zero inflation: Let us assume that we know that only a zero inflation is possible for given data. In this case, for the maximum likelihood estimation, the constraints given above for the ZIP distribution have also to be taken into consideration. Formally, the asymptotic variance obtained from the information matrix is equivalent to (16), so that the estimate of the Poisson parameter on basis of the ZTP model is asymptotical fully efficient. In the finite sample case, however, a loss of efficiency occurs. To study the extent of this loss, samples of ZIP models for the several Poisson parameters values and for the p value 0.8 have been generated. For each Poisson parameter

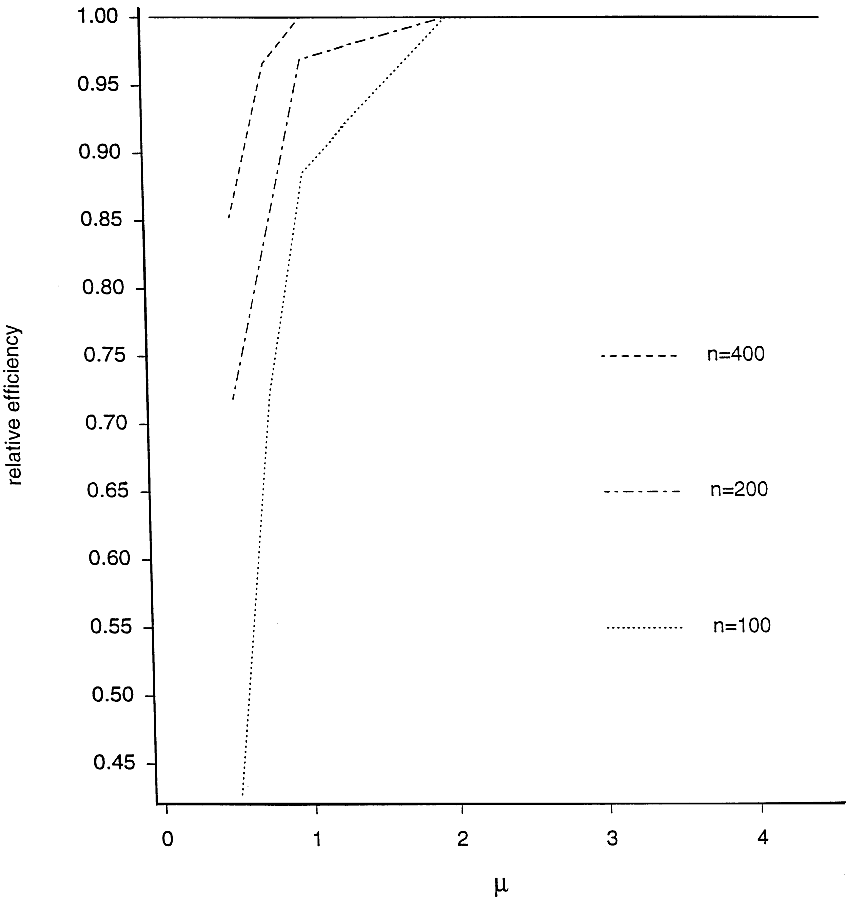

Fig. 3. Relative efficiency of the ZTP-estimation if a ZIP-model is true.

value, 1000 samples of size 100, 200 and 400, respectively, were generated. For each sample the ML estimate of the Poisson parameter was computed on basis of both the ZTP model and the ZIP model. The ML estimates of the parameters of a ZIP distribution could have been obtained by a simple EM algorithm. It was more convenient, however, to use the result of the ZTP model estimation obtained by the fixed-point iteration (13). If (15) leads to an estimate of parameter p which is smaller than one, then these ML estimates of p and μ of the ZMP model are also the ML estimates of the parameters of the ZIP model. If the estimate of p is larger than one, then the respective ML estimate for the ZIP distribution is equal to 1, so that a simple Poisson fit leads to the ML estimate of the Poisson parameter (Umbach, 1981).

The results are displayed in Fig. 3, which clearly shows the improvement of the relative efficiency of the ZTP estimate, if the sample size is increased.

In order to show the dependency of the relative efficiency on the degree of zero inflation, we simulated samples of size 100 of ZIP distributions for several values of μ and p.

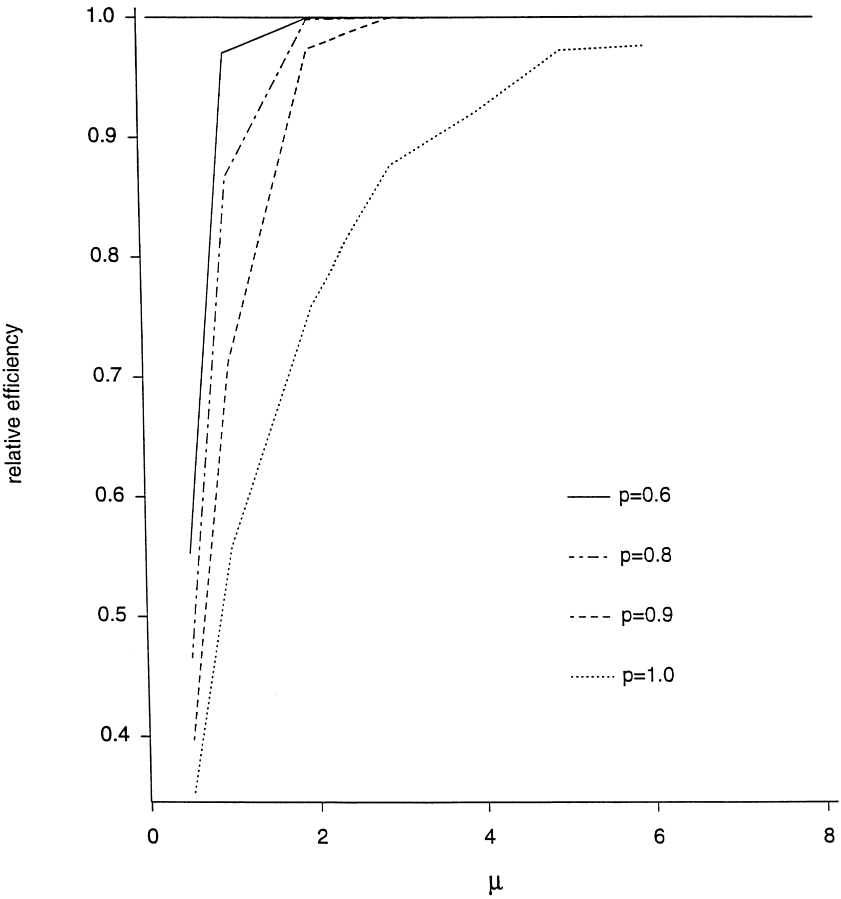

Fig. 4. Relative efficiency of the ZTP-estimation (sample size 100) if a ZIP-model is true, $\mu = 0.5, 1, 2, 3, 4, 5, 6$.

The results of the estimation of the relative efficiency based on 1000 replications for each parameter pattern are presented in Fig. 4. Fig. 5 shows the respective results, when using a two-stage sampling as described above. As can be seen in Figs. 4 and 5, the loss and the gain, respectively, of efficiency depends on the degree of zero inflation.

The loss of efficiency becomes smaller and the gain of efficiency becomes larger, respectively, if the zero inflation increases. In other words, zero inflation increases the contributed information of a positive outcome to the Poisson parameter estimate.

4.3. ZTP regression

Let n and n_0 denote the sample size and the number of sample elements having a zero response, respectively. The sample elements are assumed to be ordered by its response, so that the zero-response elements take the first positions. The log

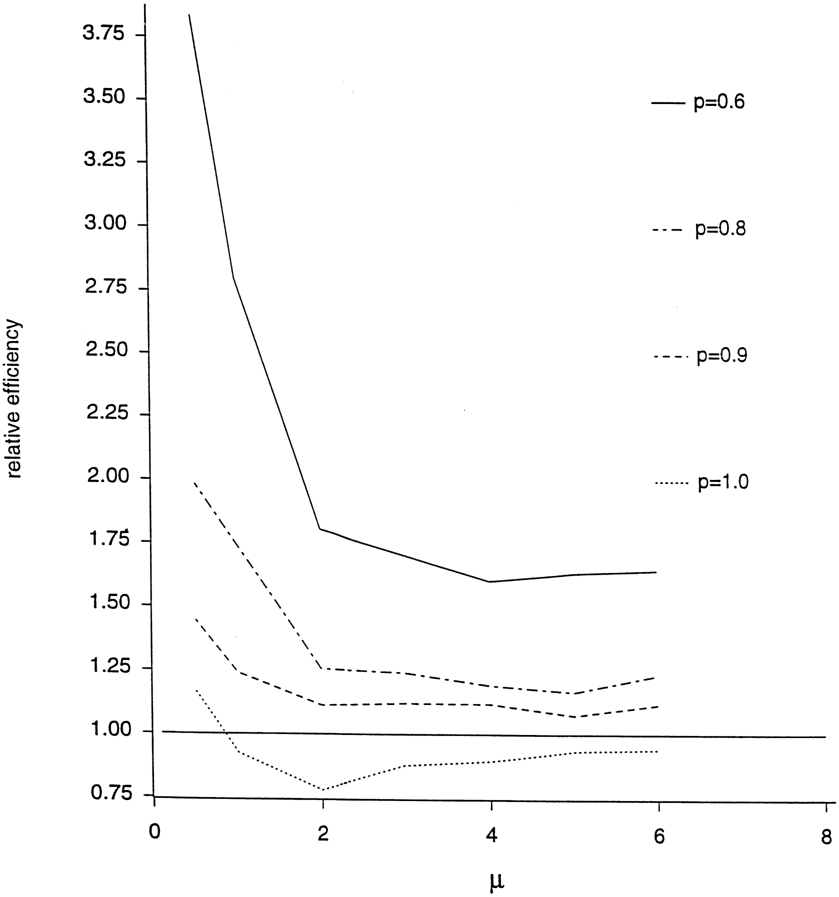

Fig. 5. Relative efficiency of the ZTP-estimation from two-step samples of size 100 if a ZIP-model is true, $\mu = 0.5, 1, 2, 3, 4, 5, 6$.

likelihood of a ZTP regression model for the positive-response sample elements is

$$LL_{ZTP} = \sum_{i=n_0+1}^{n} [y_i \log(\mu(\beta_1^{\mathrm{T}} x_i)) - \mu(\beta_1^{\mathrm{T}} x_i) - \log(1 - \mathrm{e}^{-\mu(\beta_1^{\mathrm{T}} x_i)}) - \log(y_i!)]. \quad (18)$$

Using the log link, we have

$$\mu(\beta_1^{\mathrm{T}} x_i) = \mathrm{e}^{\beta_1^{\mathrm{T}} x_i}.$$

In order to obtain the ML estimation of β_1, one has to solve the likelihood equations obtained by setting the components of the score vector

$$\begin{aligned} u(\beta_1) &= \sum_{i=n_0+1}^{n} (y_i - \mathrm{e}^{\beta_1^{\mathrm{T}} x_i}) x_i - \frac{\mathrm{e}^{\beta_1^{\mathrm{T}} x_i} \mathrm{e}^{-\mathrm{e}^{\beta_1^{\mathrm{T}} x_i}}}{1 - \mathrm{e}^{-\mathrm{e}^{\beta_1^{\mathrm{T}} x_i}}} x_i \\ &= \sum_{i=n_0+1}^{n} (y_i - \mathrm{e}^{\beta_1^{\mathrm{T}} x_i}) x_i + \sum_{i=n_0+1}^{n} w_i (0 - \mathrm{e}^{\beta_1^{\mathrm{T}} x_i}) x_i \end{aligned} \quad (19)$$

equal to zero. Eq. (19) is also the score function of a standard Poisson regression model and the data at hand, (y_i, x_i), $i = n_0 + 1, n_0 + 2, \ldots, n$, augmented by the $n - n_0$ w_i-weighted data records $(0, x_i)$. The weights

$$w_i = \frac{\mathrm{e}^{-\mathrm{e}^{\beta_1^{\mathrm{T}} x_i}}}{1 - \mathrm{e}^{-\mathrm{e}^{\beta_1^{\mathrm{T}} x_i}}} \tag{20}$$

are just the expected number of a zero response given the value of the vector of covariables, x and the parameter vector β_1. Thus, the solution of the likelihood equation can be obtained by an EM algorithm, which can be performed by standard Poisson regression software. Thereby, in the E-step, the w_i have to be computed using the current estimate of β_1. Thus, we have a multivariate generalization of the fixed point algorithm (12).

4.4. ZMP regression vs. ZTP regression

4.4.1. ML estimation of ZMP regression models

The log likelihood of a general ZMP regression model (4) is

$$LL_{ZMP} = \sum_{i=n_0+1}^{n} [y_i \log(\mu(\beta_1^{\mathrm{T}} x_i)) - \mu(\beta_1^{\mathrm{T}} x_i) + \log(p_i) - \log(y_i!)] + \sum_{i=1}^{n_0} [\log(1 - p_i(1 - \mathrm{e}^{-\mu(\beta_1, x_i)}))].$$

It holds

$$LL_{ZMP} = LL_{ZTP} + \Delta,$$

where

$$\Delta = \sum_{i=1}^{n_0} [\log(1 - p_i(1 - \mathrm{e}^{-\mu(\beta_1^{\mathrm{T}} x_i)}))] + \sum_{i=n_0+1}^{n} [\log(p_i(1 - \mathrm{e}^{-\mu(\beta_1^{\mathrm{T}} x_i)}))].$$

Because $\mu(\beta_1, x_i) > 0 \ \forall i$, Δ is maximized at

$$\hat{p}_i = \begin{cases} 0 & \text{if } y_i = 0, \\ 1/(1 - \mathrm{e}^{-\mu(\beta_1^{\mathrm{T}} x_i)}) & \text{if } y_i > 0, \end{cases} \quad i = 1, 2, \ldots, n.$$

The maximum is 0, quite independent on β_1. Thus, LL_{ZTP} is just the profile likelihood of β_1 for given $p_i = \hat{p}_i$ of the ZMP regression model and the ML estimation of β_1 of the most general ZMP regression model is identical to the ML estimation of the ZTP regression model based on the "positive" sample elements.

4.4.2. Relative efficiency

The previous section contains a proof of the following result:

Result 2. *The ML estimate of β_1 based on the ZTP regression model is fully efficient for the respective parameter of the most general ZMP regression model.*

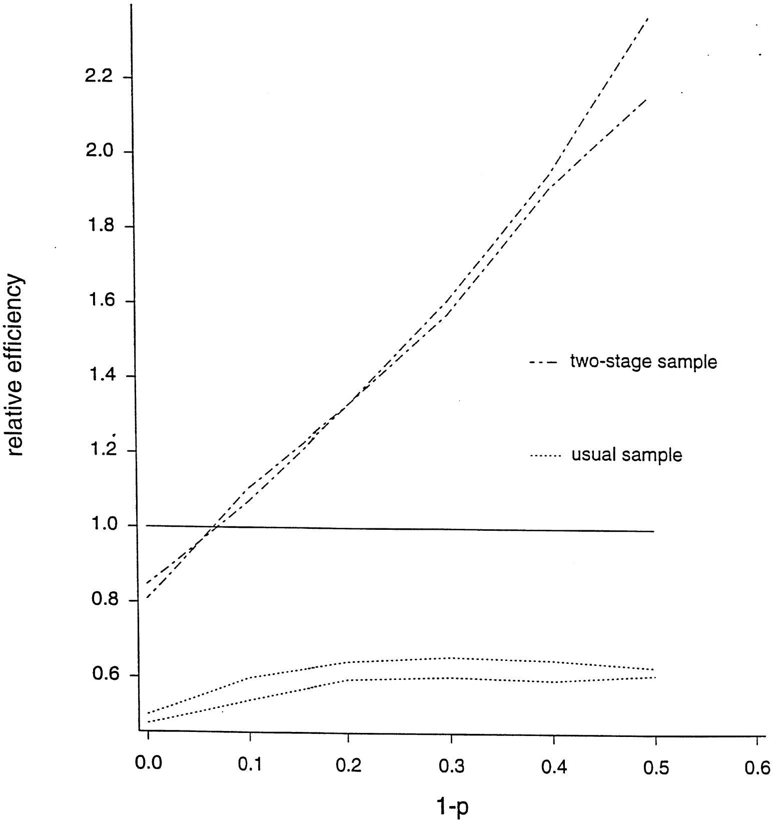

Fig. 6. Relative efficiency of the ZTP-regression vs. ZIP-regression.

The situation changes, however, if some prior information on the p_i is used. Such prior information can be introduced by a model for the p_i as described in Section 2.2.

Simulation experiments showed that the loss of efficiency is usually low. If the two-stage sampling is used an increase of efficiency can even occur. To study this more systematically, worst-case situations are considered. For example, we simulated samples of ZIP regression models having one explanatory variable x, where both the intercept and the slope parameter are equal to 1, where p is assumed to be a constant, and where x varies in a way, that $\mu(\beta_1^{\mathrm{T}}x)$ varies mainly in the most problematic region between 0 and 8. The dotted lines in Fig. 6 show the estimated relative efficiency of the slope and the intercept parameter as function of p, when using usual random sampling. They are estimated from 1000 samples of size 100. As can be seen, the relative efficiency is, even in this situation, always larger than 0.5, for each value of p considered, whereby the relative efficiency of the slope parameter estimate was about the same as those of the intercept.

The respective results for the two-stage samples are represented by the dashed lines in this figure. They show estimated relative efficiencies larger than 1 for $(1-p)$ values larger than 0.1.

Table 1
Estimates of prevention effect on the DMFT index by ZMP regression models with the additional covariables gender and ethnic group

Model	Log likelihood	PREV	p	p_{control}	p_{prev}
Poisson	−2367.0	−0.299 (0.064)	—	—	—
ZMP1[a]	−1417.5	−0.178 (0.064)	0.79	—	—
ZMP2[a]	−1415.0	−0.178 (0.064)	—	0.87	0.77
ZIP1	−1417.2	−0.217 (0.064)	0.79	—	—
ZIP2	−1414.8	−0.178 (0.064)	—	0.87	0.77

[a]Computed by the two-stage procedure.

5. Results for the example data

The first line of Table 1 contains the result of a usual Poisson regression model fit of the example data. Besides the binary study factor "PREVention", the categorical covariables gender and ethnic group (3 categories: dark(baseline) ,white(2), and black (3)) were also used in this study. The goodness of fit of this model is worse and conclusions on the prevention effect are hardly possible because of the obvious zero modification. More reliable conclusions can be drawn from the fit of two ZMP models by the two-step procedure. These two models differ by the model of the zero modification. The first model (ZMP1) assumes a fixed zero modification. The second model (ZMP2) allows the zero modification to depend on the dichotomous study factor.

Both models have a clearly improved fit and show a significant prevention effect which is quite different from those of the standard Poisson regression fit. The advantage of model ZMP2 is its power to estimate a potential other aspect of the prevention effect, which is the difference between p_{control} and p_{prev}. So, not only a decrease in the Poisson parameter can be shown but also an increase of the zero inflation in the prevention schools.

Lines 4 and 5 of Table 1 contain the ML estimates of respective ZIP models. Because the estimates of the zero inflation are always smaller than 1, they are also the ML estimates of the zero modification in the respective ZMP model. Notice, that the ML estimate of the linear predictor of the Poisson parameter depends on the model of zero modification, whereas the respective two-step estimate is independent of this model !

To demonstrate the full power of the method exposed in this paper, we analyzed additionally a somewhat modified data set. In two of the five prevention schools, we deleted randomly all but a few of those data records, which have a zero DMFT index. Practically, such school specific subsamples could occur, e.g., if, falsely, at the beginning of the study only children having a positive DMFT index are taken into the sample. In such a situation, the question, if and how the data of these two schools can be used in the statistical analysis of a common prevention effect may arise.

Fig. 7 shows the respective empirical distribution of the DMFT index in the two schools with the artificial false recruitment. This distribution shows a zero deflation

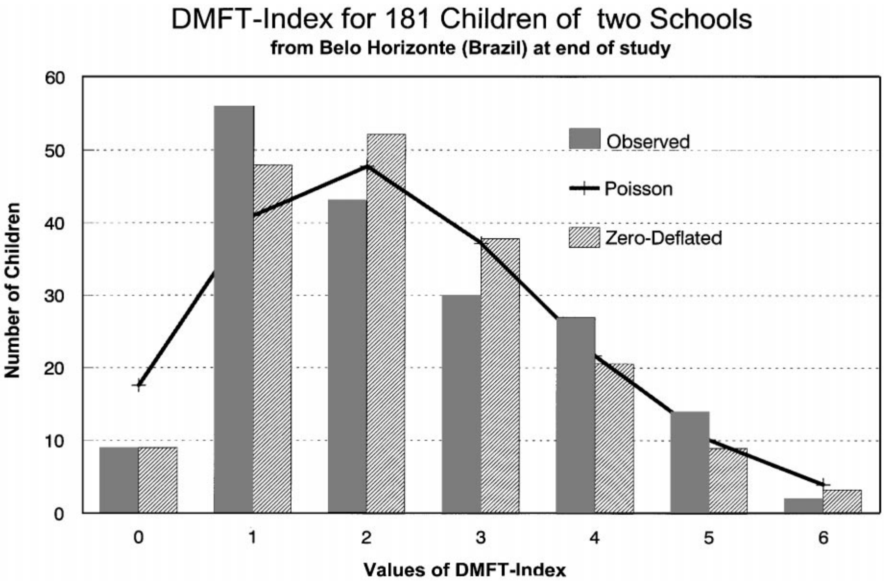

Fig. 7. Values of DMFT-index.

Table 2
Estimates of prevention effect on the DMFT index by ZMP regression models with the additional covariables gender and ethnic group from the modified data

Model	Log likelihood	PREV	p	p_{control}	p_{prev}	p_{faulty}
Poisson	−1294.3	−0.133 (0.064)	—	—	—	—
ZMP1[a]	−1275.6	−0.178 (0.067)	0.90	—	—	—
ZMP2[a]	−1275.2	−0.178 (0.067)	—	0.87	0.91	—
ZMP3[a]	−1256.9	−0.178 (0.067)	—	0.87	0.83	1.06
ZIP1	−1275.5	−0.153 (0.067)	0.90	—	—	—
ZIP2	−1275.1	−0.175 (0.067)	—	0.87	0.91	—
ZIP3	−1258.1	−0.151 (0.067)	—	0.87	0.83	1.00

[a] Computed by the two-stage procedure.

if compared with the standard Poisson distribution. Table 2 contains the respective model fits for this data. Additionally, a third model (ZMP3) is considered. This model allows the zero modification to depend on a categorical factor with the three categories “control school”, “prevention school with correct sampling”, and “prevention school with incorrect sampling”.

It leads to further improvement of goodness of fit and detects the zero deflation in the two prevention schools with the incorrect sampling procedure. The respective ZIP model (ZIP3) is not able to detect this zero deflation. As a consequence, the ML-estimated prevention effect of this model, which is no longer the ML estimate of the ZMP model, is biased.

6. Conclusions and open questions

From the results for the example data, it is evident that a wrong model for the zero modification in a ZMP model can result in a biased estimate of the Poisson parameter or its predictor, respectively. In many applications, reliable information on the kind of zero modification is not available. In such cases, the ZTP approach seems to be the most appropriate one. If nothing is known about the kind of zero modification, the ZTP approach has shown to be fully efficient, with respect to the Poisson parameter and its predictor, respectively. The ZTP estimate is also fully efficient if certain kinds of zero modification can be assumed. This holds, if the link function 5 for the modification parameter is true. Specifically this holds, if a constant zero modification parameter is assumed and the Poisson parameter does not depend on covariables.

In many other situations, where prior information on the kind of zero modification is available, the loss of efficiency turns out to be small. This is especially the case, if the Poisson parameter value(s) is (are) larger than 8.

If the two-stage sampling procedure is used, even a gain in efficiency is possible.

The estimation of a ZTP regression model can be easily done by standard Poisson regression software as shown in Section 4.

Some interesting questions with respect to the ZTP approach could not be included in this paper and are left to future research. Three of these are:

The estimation of the zero modification parameter p: In some application, the parameter p rather than the Poisson parameter μ is of special interest. For the ML estimates of p, one needs the zero observations of the sample. One could use the conditional ML estimate of p given the ML estimate of μ by the ZTP approach. In special situations, also this estimator is fully efficient. For ZMP distributions, e.g., Eq. (15) provides both the conditional ML estimate and the unconditional ML estimate of p. It would be interesting to see how efficient such a procedure can be in more general situations.

A likelihood ratio test of zero modification: The ZTP approach can be used to compute a likelihood-ratio test of zero modification. If one considers zero modification with a non-varying parameter p, this test turns out to be the generalized likelihood ratio test for zero inflation (and zero deflation) given in Feng and McCulloch (1992). Simulations show that the distribution of this statistic is very well approximated by a chi square distribution with one degree of freedom, even for small sample sizes like 100. For the most general zero modifications, the distribution of the LR statistic depends on the sample size. The use of Monte-Carlo-methods should be studied.

Finite mixtures of ZMP models: In some applications, a mixture of ZMP models rather than a simple ZMP model may be appropriate. Also in this case, the ZTP approach seems to be useful, because it is easy to see that a zero-truncated mixture of ZMP models is a mixture of ZTP models having the same Poisson parameters. The mixture weights are, however, different. They are dependent not only on the original mixture weights but also on the Poisson parameters and the zero-modification parameters of the mixture components. The possibility of estimating the original weights from the ZTP-mixture fit have to be investigated.

References

Cohen, A.C., 1960. An extension of a truncated Poisson distribution. Biometrics 16, 447–450.

David, F.N., Johnson, N.I., 1952. The truncated Poisson. Biometrics 8, 275–285.

Feng, Z., McCulloch, C.E., 1992. Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. Statist. Probab. Lett. 13, 235–332.

Gurmu, S., 1991. Tests for detecting overdispersion in the positive Poisson model. J. Bus. Econom. Statist. 9, 215–222.

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

McKendrick, A.G., 1926. Applications of mathematics to medical problems. Proc. Edinburg Math. Soc. 44, 98–103.

Mendonca, L., 1995. Longitudinalstudie zu kariespräventiven Methoden, durchgeführt bei 7- bis 10-jährigen urbanen Kindern in Belo Horizonte(Brasilien). Inaugural-Dissertation zur Erlangung der zahnmedizinischen Doktorwürde am Fachbereich Zahn-, Mund- und Kieferheilkunde der Freien Universität Berlin.

Umbach, D., 1981. On inference for a mixture of a Poisson and a degenerate distribution, Comm. Statist. – Theory Methods, A 10, 299–306.

Xie, T., Aickin, M., 1997. A truncated Poisson regression model with application to occurrence of adenomatous polyps. Statist. Med. 16, 1845–1857.