Original Paper

Caries Research

Caries Res 2016;50:517–526
DOI: 10.1159/000448197

Received: August 20, 2015
Accepted after revision: July 5, 2016
Published online: September 17, 2016

Modeling Caries Experience: Advantages of the Use of the Hurdle Model

Hedwig Hofstetter[a] Elise Dusseldorp[a, b] Achim Zeileis[d]
Annemarie A. Schuller[a, c]

[a]TNO (Netherlands Organization for Applied Scientific Research), Expertise Group Life Style, and [b]Methodology and Statistics, Institute of Psychology, Leiden University, Leiden, and [c]Centre of Dentistry and Oral Hygiene, University Medical Centre Groningen, Groningen, The Netherlands; [d]Department of Statistics, Faculty of Economics and Statistics, University of Innsbruck, Innsbruck, Austria

Key Words
Count data · Dental caries · Excess zeros · Hurdle model · Overdispersion · Zero-inflated model

Abstract
In dental epidemiology, the decayed (D), missing (M), and filled (F) teeth or surfaces index (DFM index) is a frequently used measure. The DMF index is characterized by a strongly positive skewed distribution with a large stack of zero counts for those individuals without caries experience. Therefore, standard generalized linear models often lead to a poor fit. The hurdle regression model is a highly suitable class to model a DMF index, but its use is subordinated. We aim to overcome the gap between the suitability of the hurdle model to fit DMF indices and the frequency of its use in caries research. A theoretical introduction to the hurdle model is provided, and an extensive comparison with the zero-inflated model is given. Using an illustrative data example, both types of models are compared, with a special focus on interpretation of their parameters. Accompanying R code and example data are provided as online supplementary material.

In dental epidemiology, the DMF index is an often-used measure that represents the number of decayed (D), missing (M), and filled (F) teeth or surfaces[1] in an individual. At least in Western countries, DMF is distributed as a count variable with specific features; it has a *strongly* positively skewed distribution with a large stack of zero counts for those individuals without caries experience. Figure 1 shows an example of such a DMF distribution. For more illustrations of typical DMF distributions, we refer to previously published papers [Böhning et al., 1999; Lewsey and Thomson, 2004; Preisser et al., 2012]. A challenge for the researcher is to assume a suitable probability distribution to model DMF. At times researchers recode the DMF index into zero versus larger than zero and use logistic regression to analyze this binary variable [Liu et al., 2013; Nobile et al., 2014]. Nevertheless, information about the amount of caries is lost [Preisser et al., 2014]. Regression models belonging to the family of generalized

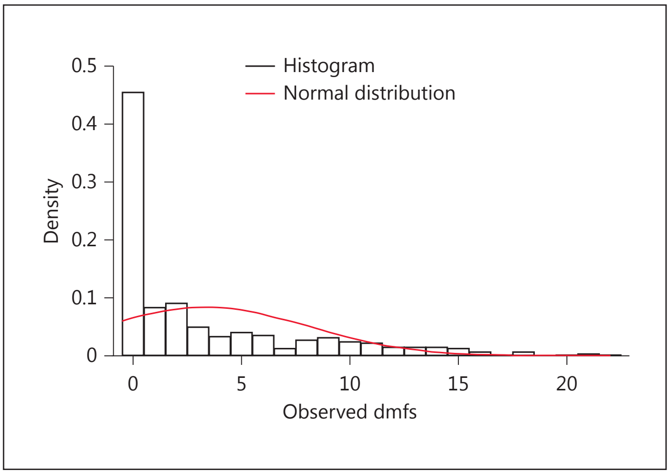
[1] The DMF index is used to describe permanent dentition, where dmf is used for deciduous teeth. When the index is used to describe teeth, it is called the DMFT (or dmft) index. For tooth surfaces, the DMFS (or dmfs) index is used.

KARGER

© 2016 S. Karger AG, Basel
0008–6568/16/0506–0517$39.50/0

E-Mail karger@karger.com
www.karger.com/cre

Hedwig Hofstetter
TNO, Expertise Group Life Style
P.O. Box 3005
NL–2301 DA Leiden (The Netherlands)
E-Mail hedwig.hofstetter@tno.nl



Color version available online

Fig. 1. Histogram of the dmfs index.

linear models, among which linear regression, Poisson regression, and negative binomial regression, often lead to a poor fit. A short overview of these aforementioned models, along with their advantages and disadvantages, is presented in table 1. These models will be further discussed in the following section.

Two closely related classes have been proposed as approaches to model the DMF distribution: the hurdle regression model [Mullahy, 1986] and the zero-inflated regression model [Lambert, 1992]. Both classes simultaneously model the DMF variable in two separate parts: a logistic part for the (excess) zero values in the data and a count regression part. They, however, take a different approach in modelling the zero values, which will be explained in this paper. In addition, various model specifications are possible within both classes. Previous studies compared the overall performance of count regression models (Poisson, negative binomial, and their zero-inflated and hurdle variants) in modeling an outcome variable with extra zeros, for example, the number of unprotected sexual occasions [Hu et al., 2011] and the number of episodes of a treatment side effect [Min and Agresti, 2005]. In this paper, we focus on modeling DMF and, more specifically, on the interpretation of the model coefficients.

In the past few years, the use of zero-inflated models has emerged in caries research [Böhning et al., 1999; Lewsey and Thomson, 2004; Solinas et al., 2009; Javali and Pandit, 2010; Campus et al., 2011]. For example, Böhning et al. [1999] have used the zero-inflated regression model for the DMFT index. A zero-inflated regression model was also used by Lewsey and Thompson [2004] to model several DMF indices. As Preisser et al. [2012] already mentioned in their review on the use of zero-inflated models in dental caries, hurdle regression models are only occasionally used in caries research [Jahani et al., 2013; Dusseldorp et al., 2015]. However, Preisser et al. [2012] show that interpretations of zero-inflated models are often incorrect because they are based on language that suits the hurdle model better. To summarize, there appears to be a gap between the suitability of the hurdle model to fit DMF indices and the frequency of its use in caries research. The present study aims to lessen this gap by explicating the use of hurdle models to researchers faced with caries indices. A theoretical introduction to the hurdle model within the framework of count regression models will be provided. In addition, an illustrative data example will be given in which the hurdle model and zero-inflated model will be compared using the same data set. Accompanying R code and example data are provided as online supplementary material (for all online suppl. material, see www.karger.com/doi/10.1159/000448197).

Count Regression Models

A typical count variable is not normally distributed and includes many zero observations. Count variables will therefore rarely meet the distributional assumptions of ordinary least squares linear regression. Although a very common strategy, transforming the outcome variable will not smooth out the large stack of zeros in the data. Count regression approaches, as their name implies, are much more appropriate techniques. The basic distribution for a count variable is a Poisson distribution, which has the characteristic that the conditional mean (the mean of the outcome variable *Y* given the values of the predictor variables *X*) should equal the conditional variance. Most count data are, however, characterized by overdispersion, that is, the variance exceeds the mean. The negative binomial distribution extends the Poisson model by allowing the mean and variance to be different by introducing an unobserved heterogeneity term *theta (θ).* The negative binomial distribution looks like the Poisson distribution, but with a longer, fatter tail to the extent that the variance exceeds the mean. Depending on the degree of overdispersion, the negative binomial model can capture (much) more zeros than the Poisson model. However, the model may still be insufficient in many empirical applications with a clear stack of zero values in the data.

Table 1. Overview of possible (not always optimal) strategies to model DMF data

Technique	Outcome variable	How?	Advantages	Disadvantages
Binomial logistic regression	Dichotomous variable	Collapse the DMF variable into two groups: recode all positive counts into 1 ('those with caries experience'), and compare them with individuals without caries experience	Relatively easy method to adapt (to take care of the large peak of zeros in the data); generally known among researchers; applicable in standard software	Loss of information; no information available about the amount of caries experience
Linear regression analysis	Continuous variable	Use the original DMF variable as outcome	Applicable in standard software	Often violates the assumption of normality; wrong conclusions in the end; mean is misspecified and can become negative
Linear regression analysis with transformed outcome variable	Transformed continuous variable	Because of the skewed distribution, transform the DMF variable and use as outcome in ordinary linear regression	Often better fit than linear regression analysis with original outcome variable	No transformation will smooth out the large stack of zeros in the data, therefore assumptions of regression analysis will still often be violated; in addition, transformation makes interpretation somewhat harder
Poisson regression	Nonnegative variable, with its mean equal to variance	Regression model in which outcome variable is assumed to follow a Poisson distribution	For nonnegative count outcomes, a model with Poisson distribution is much more appropriate than linear regression	Assumes that mean and variance are equal, which is often not the case for count data (overdispersion); as a result, standard errors will be biased downwards leading to incorrect statistical inferences; observed number of zeros might exceed the estimated zero values
Negative binomial regression	Variable where variance exceeds the mean	Generalization of Poisson regression model with extra parameter to model overdispersion	More appropriate for overdispersed count data than Poisson (variance exceeds the mean)	Not always suitable for data with large stack of zeros; can capture more zeros than the Poisson model, but potentially still not enough

Count Data Models with Excess Zeros

Zero-inflated models and hurdle models provide a way of modeling the excessive proportion of zero values and allow for overdispersion. Especially when there is a large number of zeros, these techniques are much better able to provide a good fit than Poisson or negative binomial models. The major difference between zero-inflated and hurdle models is in the way the zero values are modeled.

Hurdle Models

A hurdle model has the following parts: one *zero hurdle part* which models a right-censored outcome variable indicating persons without ($Y = 0$) or with caries experience ($Y = 1$, where all values larger than 0 are censored, that is, are fixed at 1), and one truncated *count part* that models the amount of caries experience for those with caries experience (for those with $Y > 0$). If an individual is without caries experience, the threshold (the 'hurdle') to the truncated count part is not crossed, and a zero value is observed. Otherwise, the hurdle to the truncated count part is crossed, and the amount of caries experience is observed.

Mathematical explanation. The hurdle model can be expressed by

$$P\left(Y_i = y_i | x_i, z_i, \beta, \gamma\right) = \begin{cases} f_{zero}\left(0; z_i; \gamma\right), & if \;\; y_i = 0 \\ \left(1 - f_{zero}\left(0; z_i; \gamma\right)\right) \dfrac{f_{count}\left(y_i; x_i; \beta\right)}{1 - f_{count}\left(0; x_i; \beta\right)} & if \;\; y_i > 0 \end{cases} \quad (1)$$

In this equation, y_i is the value of the dependent variable for the i^{th} person $i = 1, …, N$), z_i is a vector of length J denoting the number of predictor variables in the zero part, χ_i represents a vector of length K denoting the number of predictor variables in the hurdle part, γ is a vector of coefficients belonging to z, and β denotes a vector of coefficients related to x [Zeileis et al., 2008]. f_{zero} is a probability density function at least on {0, 1} (binary) or {0, 1, 2, …} (count), and f_{count} is a probability density function on {0, 1, 2, …}. Different distributions can be plugged in with some common choices presented in more detail below. Regression coefficients are estimated with maximum likelihood.

The f_{zero} part in equation 1 (where $y_i = 0$) is typically modeled with a binary logit (logistic regression) model, where all counts greater than 0 are given a value of 1. In this zero part, *all* zeros are estimated, as subjects with no caries are supposed to come from a single population. Using a binary logistic regression model for this part, the probability of $y_i = 0$ is denoted as

$$f_{zero}\left(0; z_i; \gamma\right) = \psi_i = \frac{1}{1+e^{z_i \gamma}} \qquad (2)$$

where z_i represents the observed data and γ the vector of coefficients belonging to z_i. Obviously, the probability of a nonzero count is given by $1 - \psi_i$.

The lower part in equation 1 *(f_{count})* is modeled with a left-truncated ($y_i > 0$) count model. This is typically a (left-truncated) Poisson model, or a negative binomial model in case of overdispersion. The Poisson model is represented as

$$f_{count}\left(y_i; x_i; \beta\right) = \frac{e^{-\mu_i}\,\mu_i^{y_i}}{y_i!} \qquad (3)$$

and is inflated by $(1/(1 - e^{-\mu_i}))$ to allow for the truncation. The μ_i represents the mean of the untruncated distribution, which is assumed to depend on observed data χ_i and β:

$$\mu_i = e^{\chi_i \beta} \qquad (4)$$

In the hurdle model, β in equation 4 is only estimated using the $y_i > 0$ observations (and is only used to model the $y_i > 0$ observations). The mean of the truncated distribution *(τ_i)* can be computed as

$$\tau_i = \frac{\mu_i}{1 - f_{count}\left(0; x_i; \beta\right)} \qquad (5)$$

Note that the difference between τ_i and μ_i vanishes as μ_i moves away from the truncation point.

The Poisson model assumes that the conditional mean is equal to the conditional variance. A less strict model is the negative binomial model, which allows the conditional mean and variance to be different. For a mathematical description of this latter model, the reader is referred to, for example, Cheung [2002] or Javali and Pandit [2010].

Because the zero and count parts are clearly separated, they can be specified and estimated separately. The model allows for a set of predictors for the probability of a zero response *(z_i)* and a different set for the mean of nonzero responses *(x_i)*. Note that the choice of the model for the count part, for example a Poisson or negative binomial, defines that we can speak of, for example, a hurdle Poisson model or a hurdle negative binomial model.

The zero hurdle part targets the probability of a person being caries free and therefore relates to prevalence in the population (the proportion of the population that has no caries experience). The odds ratios that are obtained in this part of the model can be directly linked to the observed probabilities of having no caries. The count part models the mean degree of caries (for those with caries experience). The rate ratios from this part of the model denote the rate ratios of untruncated means (see equation 4). In other words, the rate ratio is the ratio of two incidence rates and represents the percentage of increase or decrease in the outcome variable for a 1-unit increase in the predictor variable (for example, comparing females with males).

Zero-Inflated Models

The zero-inflated model assumes that the zero counts arise from a mixture of two sources, that is, from a zero component distribution (to model the excess zeros) and from a count distribution (e.g. Poisson or negative binomial). In other words, a zero count can be estimated in either the zero part or in the count distribution part. A graphical example of a (typical zero-inflated) distribution represented as a mixture of two sources of zeros is given in, for example, Preisser et al. [2012]. The zero part represents the subpopulation of individuals who are considered not to be at risk for dental caries, whereas the count part represents those individuals who are at risk for dental caries. The difference between these two groups of individuals is a theoretical one; group membership of an individual with a zero count in the sample is unknown. This is in sharp contrast to hurdle models, where zero values can only be modeled in the hurdle part. Because of this mixture modeling of zeros, one has to be careful with inferences for the overall population. In fact, in some studies the coefficients of a zero-inflated model were 'wrongly' interpreted or even interpreted as a 'hurdle' model, something that has been greatly expounded in a review by Preisser et al. [2012] (see table 2). Severity in the overall population is determined with the overall incidence rate ratio (IRR), that is, the ratio of overall means.

Mathematical explanation. The zero-inflated model is expressed by

$$P\left(Y_i = y_i \mid x_i,, \psi_i^*, \beta\right) = \begin{cases} \psi_i^* + \left(1-\psi_i^*\right) f_{count}\left(0; x_i; \beta\right) & \text{if } y_i = 0 \\ \left(1-\psi_i^*\right) f_{count}\left(y_i; x_i; \beta\right) & \text{if } y_i > 0 \end{cases} \qquad (6)$$

Table 2. Evaluating goodness of fit of the hurdle models and zero-inflated models applied to example data

	Hurdle Poisson	Hurdle negative binomial	Comparison	Zero-inflated Poisson	Zero-inflated negative binomial	Comparison
log likelihood	–968.76	–838.24		–978.79	–839.48	
AIC	1,969.53	1,710.48		1,969.57	1,712.95	
BIC	2,033.23	1,778.16		2,033.27	1,780.64	
χ^2, d.f.[a]			261.05 (1)**			258.62 (1)**

AIC = Akaike information criterion; BIC = Bayesian information criterion; d.f. = degrees of freedom. ** p < 0.01.

[a] With the χ^2 test, the hurdle Poisson is compared with the hurdle negative binomial model, and the zero-inflated Poisson with the zero-inflated negative binomial.

where y_i is the value of the dependent variable for the i^{th} person (i = 1, …, N), ψ_i* is the probability of an extra zero, χ_i is a vector of length K denoting the number of predictor variables in the count part, and β is a vector of coefficients related to χ. f_{zero} is a probability density function at least on {0, 1}, and *f(count)* is a probability density function on {0, 1, 2, …}. Different distribution can be used, with some common choices represented below. Regression coefficients are estimated with maximum likelihood.

Unlike in the hurdle model where *all* zeros are modeled in f_{zero}, in the zero-inflated model only *excess* zeros are estimated in the f_{zero} part. More specifically, individuals with y_i = 0 can be part of two groups: one group (the excess zeros, or the ‘not at risk’ group) that is not part of the count process, and one group (the sampling zeros, or the ‘at risk’ group) that belong to the count (e.g. Poisson or negative binomial) distribution with mean μ but only taking zero values. For y_i >0, all are considered part of the count distribution.

Following Preisser et al. [2012], the estimated probability of observing an *excess* zero is given by

$$f_{zero}\left(0; z_i; y\right) = \psi_i^* = \frac{e^{z_i y}}{1 + e^{z_i y}} \qquad (7)$$

where z_i is a vector of length J denoting the number of predictor variables in the zero part, and y is a vector of coefficients related to z_i.

The lower part in equation 6, that is, *f(count)*, is typically modeled with a Poisson (see equation 3) or negative binomial model. In contrast to the hurdle model, β is estimated from all Y observations (Y ≥0).

As in hurdle models, a different set of predictor variables can be used in the zero and count parts. Also, the count part can be modeled by a Poisson or negative binomial model. Note that the choice of the distribution implies that we can speak of, for example, a zero-inflated Poisson or a zero-inflated negative binomial model.

Choice between Hurdle and Zero-Inflated Models

The choice between a hurdle model and a zero-inflated model should to an extent be based on theoretical considerations. A zero-inflated model assumes that the zero values can have two different sources. The subpopulation of individuals who are ‘not at risk’ for caries is modeled in the zero part, whereas the subpopulation of individuals who are ‘at risk’ is modeled in the count part. In this latter part, it is still possible to observe zero values (due to the usual Poisson or negative binomial distribution). Therefore, the zero part estimates *excess* zeros, that is, above and beyond the zero values modeled in the count part. In contrast, a hurdle model assumes that all zero values come from the zero hurdle part, and therefore the positive counts are modeled with a truncated distribution. The two parts of the hurdle model are separable in terms of parameters being estimated (using likelihoods) and hence can be fit in two separate steps. A zero-inflated model assumes the existence of both excess zeros (the individuals who are ‘not at risk’) and sampling zeros (the individuals who are ‘at risk’), and the likelihoods with respect to parameters being estimated cannot be separated. For this reason, a hurdle model is easier to implement and to interpret. The theoretical difference between hurdle and zero-inflated models may be subtle, but statistical decisions can support the choice between the two techniques. The Vuong statistic [Vuong, 1989] can be computed to test nonnested models and can therefore be used to test the validity of the hurdle model against the zero-inflated model. A nonsignificant Vuong statistic indicates that both models are equally close to the true model, while a significant Vuong model suggests that one of the models is closer to the true model. In addition, information criteria, such as Akaike information criterion [Akaike, 1973] and Bayesian information criterion [Schwarz, 1978], can be used to assess the fit of the models. Besides these formal tests, one can examine how well the models estimate the number of zeros in the observed data. Furthermore, visualizations are a handy tool to explore the (goodness of) fit of the models. For example, rootograms can be used to compare the observed and fitted frequencies [Kleiber and Zeileis, 2014]. However, differences between the hurdle and zero-inflated model are often small (see, for instance, the results in the practical example presented below), hence the choice between the models can also be easily made based on ease of interpretation.

Hurdle and zero-inflated models are both able to cope with a large proportion of zero values in an outcome variable. For modeling the *count part* in these models, a choice has to be made for the best-fitting distribution. The most common approach is to assume either a Poisson or a negative binomial distribution for the count part of the model. A leading aspect in the choice between these two distributions is the amount of overdispersion. By examining the mean-variance ratio, we can get an indication whether overdispersion is present. Another indication of overdispersion is given by the difference in log likelihood values of the Poisson and negative binomial distribution within a hurdle model or within a zero-inflated model. Formal tests are available that can guide the researcher in choosing the model with the best fit: since the negative binomial distribution extends the Poisson model, these two models are so-called nested models, and their fit can therefore be tested with the *likelihood ratio statistic*. The difference in log likelihoods of a model with a Poisson versus a model with a negative binomial distribution is tested against a χ^2 distribution, with degrees of freedom equal to the difference in parameters between the models. The p value from this χ^2 distribution should be divided in half (test the difference in log likelihoods against a p value of 0.1) for a correct significance level[2]. For the choice between the Poisson hurdle and the negative binomial hurdle model, the value of θ can also be inspected. The symbol θ is a dispersion parameter, and the value of θ would be constrained to be infinite in a Poisson model. In the case where θ is less than infinite, a negative binomial model would give a better fit. If in doubt, however, it is recommended to use the negative binomial distribution instead of the Poisson. If a negative binomial is used while the 'true' model is Poisson, only a little bit of efficiency is lost by estimating one parameter too many (for the θ parameter). By contrast, if a Poisson is estimated while the 'true' model is a negative binomial distribution, the model is misspecified. As a result, coefficient estimates will be consistent, but standard errors will be too small, leading to liberal significance tests.

Practical Example

Data

To illustrate the use of the hurdle model, we employ data from the study 'Oral health in children and adolescents in the Netherlands' [Schuller et al., 2011]. The aim of this study was to describe the oral health status and the preventive dental behaviors of children from different age groups. The data collection consisted of a clinical oral examination and a questionnaire survey, using a repeated cross-sectional design. Data contained information about demographic variables (ethnicity and educational level), nutrition, children's dental attendance, oral self-care, and dental anxiety. The clinical assessment consisted of visual inspection of the teeth and the registration of caries lesions and any subsequent treatment (restoration or extraction). We used data from children at the age of 9 years. The same data set was also used in a paper by Dusseldorp et al. [2015]. However, in contrast to Dusseldorp et al. [2015], who focused on lifestyle factors, we also used information about dental anxiety for illustrative purposes. One outlier was removed before the analyses. All analyses were performed in the R software environment [R Core Team, 2013]. The data set and computer code to run the analyses in R are available as online supplementary material.

The dmfs score was used to describe caries experience in 9-year-olds, where the lowercase abbreviation refers to deciduous teeth. The education level of the mother was based on the highest level of completed education, with senior general secondary education (HAVO) or higher classified as 'high education' and all other as 'low education'. The gender of the child was defined as being a boy or a girl. Ethnicity was defined as the mother being born in the Netherlands (referred to as natives) versus being born abroad (referred to as immigrants). Frequency of brushing teeth was dichotomized according to Dutch norms into 'less than twice a day' versus 'at least twice a day'. Frequency of having breakfast was dichotomized

[2] The likelihood ratio test to choose between Poisson and negative binomial is not fully standard, because the null hypothesis of a Poisson model corresponds to a parameter value for θ on the boundary of the parameter space ($\theta = \infty$). The reader is referred to Molenberghs and Verbeke [2012] for more (technical) detailed information.

into 'not daily' versus 'daily', as was used in Dusseldorp et al. [2015]. Frequency of food and drinks per day (in addition to the three main meals) was classified according to Dutch norms as 'maximum 7 times daily' and 'more than 7 times daily'. The score on Corah's Dental Anxiety Questionnaire was used as a measure of dental anxiety. This questionnaire consists of four questions with answer categories from 1 (low anxiety) to 5 (high anxiety). A total Corah score was computed by taking the sum of the four items. The Corah score was dichotomized into 'lower than 13' and 'higher than or equal to 13'.[3]

Both the hurdle model and the zero-inflated model were applied to the data to assess the impact of demographic variables, lifestyle factors, and dental anxiety on the dmfs score. Analyses were based on 396 children who had complete data on these variables.

Exploratory analysis (and results from Dusseldorp et al. [2015]) already suggested that overdispersion was present in the data (mean = 3.5, standard deviation = 5.1). The distribution of the dmfs score is depicted in figure 1 and shows a clearly nonnormal distribution with an enormous stack of zero values. Therefore, we fitted a hurdle and a zero-inflated model to analyze the dmfs score. For the count part of both models, we fitted both a Poisson and a negative binomial distribution. Using a likelihood ratio test, we assessed which distribution fitted best. In the case of a significant likelihood ratio test, the restricted model with the lower (log) likelihood is rejected and the unrestricted model with the higher (log) likelihood is chosen. With a nonsignificant test, both models perform equally well, and the most restricted model is preferred. To determine whether the hurdle or the zero-inflated model fitted best, a Vuong test was performed. Analyses were performed using the hurdle() and zeroinfl() functions from the pscl package [Zeileis et al., 2008]. The same set of predictor variables was used for the zero and the count parts.

Results

The distribution of the dmfs index is shown in figure 1. In table 2, the fit of the four models (that is, hurdle Poisson, hurdle negative binomial, zero-inflated Poisson, and zero-inflated negative binomial) are shown. The log likelihood values indicate that the count part of the hurdle model (positive counts) is better modeled with a negative binomial distribution than with a Poisson distribution: the χ^2 statistic is significant ($\chi^2(1) = 261.05$, $p < 0.01$), and the hurdle model with a negative binomial distribution shows the lowest log likelihood value. The same is also true for the zero-inflated model: the values of the log likelihood and the χ^2 statistic indicated that the negative binomial (likelihood ratio = –839.48) is a better fit than the Poisson (likelihood ratio = –979.79). Comparing the hurdle model with the zero-inflated model, the nonsignificant Vuong statistic (Vuong = –2.51, $p > 0.05$) indicated that the negative binomial hurdle model and the zero-inflated negative binomial model showed comparable fit. In addition, both models predicted nearly the same number of zero values, which was (practically) equal to the number of observed values in the data (n = 176).

In tables 3 and 4 parameter estimates of the negative binomial hurdle model and the zero-inflated negative binomial model are shown, respectively. As can be seen from these tables, coefficients from the negative binomial part do not differ much. Note the opposite signs for the coefficients of the variables in the zero part from both models. Whereas in a hurdle model the zero part represents the probability of observing a positive count (dmfs >0), in a zero-inflated model this part estimates the probability of observing an excess zero. Note, however, that besides the opposite signs, the coefficients for the estimated zeros (hurdle model) and the excess zeros (zero-inflated model) do not differ much in this particular example. This does not always hold for other data sets.

The interpretation of the coefficients from the hurdle model is relatively straightforward. The zero hurdle part is a binary logistic regression (see also table 1), therefore most researchers will move on familiar ground here. For example, the odds ratio for having breakfast equals exp(1.26) = 3.52 (adjusted for the other variables), which implies that those not having breakfast every day are 3.5 times more likely to have caries experience than those having breakfast every day. In other words, the odds of observing no caries experience is significantly higher for children who take a daily meal in the morning. Although having breakfast contributes to having no caries experience, this predictor plays no significant role in modeling the amount of caries experience (count part). However, brushing teeth less than twice a day contributes significantly to the amount of caries; from the count part we find that the adjusted rate ratio for brushing teeth is exp(0.38) = 1.46. Note the huge coefficient for the Corah score in the zero part due to the quasi-complete dichotomization; all children with a Corah score of 13 or greater

[3] In this example only dichotomous predictor variables were used. Of course, one can also include continuous predictor variables [Zeileis et al., 2008].

Table 3. Estimated coefficients, odds ratios, rate ratios, and log likelihood value for the negative binomial hurdle model

	Coefficient zero part	OR (95% CI)	Coefficient hurdle part	RR (95% CI)
Intercept	–0.33	0.72 (0.49–1.06)	1.29**	3.64 (2.82–4.71)
Education level of mother (low)	0.40	1.50 (0.99–2.28)	0.30*	1.36 (1.05–1.75)
Gender (male)	–0.04	0.96 (0.63–1.45)	0.05	1.05 (0.82–1.34)
Ethnicity (immigrant)	0.49	1.63 (0.92–2.88)	0.34*	1.40 (1.04–1.88)
Brushing teeth (<2 a day)	0.26	1.30 (0.77–2.20)	0.38**	1.46 (1.10–1.93)
Having breakfast (not every day)	1.26**	3.52 (1.38–8.94)	0.14	1.15 (0.80–1.65)
Food and drinks (>7 times a day)	0.98*	2.67 (1.09–6.54)	–0.08	0.92 (0.63–1.35)
Corah score (≥13)	16.15	Inf (0.00–inf)	0.38	1.47 (0.95–2.26)
log θ	0.51**			
log likelihood, d.f.	–838.24 (17)			

CI = Confidence interval; OR = odds ratio; RR = rate ratio; d.f. = degrees of freedom. * p < 0.05, ** p < 0.01.

Table 4. Estimated coefficients, odds ratios, incidence rate ratios, and log likelihood value for the negative binomial zero-inflated model

	Coefficient zero part	OR (95% CI)	Coefficient count part	IRR (95% CI)
Intercept	0.06	1.07 (0.67–1.69)	1.31**	3.69 (2.88–4.74)
Education level of mother (low)	–0.34	0.71 (0.43–1.16)	0.30*	1.35 (1.05–1.74)
Gender (male)	0.07	1.07 (0.66–1.74)	0.05	1.05 (0.83–1.34)
Ethnicity (immigrant)	–0.46	0.63 (0.33–1.22)	0.33*	1.38 (1.04–1.85)
Brushing teeth (<2 a day)	–0.24	0.78 (0.43–1.44)	0.35*	1.42 (1.08–1.88)
Having breakfast (not every day)	–1.41*	0.25 (0.07–0.88)	0.15	1.16 (0.81–1.65)
Food and drinks (>7 times a day)	–1.21	0.30 (0.08–1.08)	–0.07	0.93 (0.64–1.36)
Corah score (≥13)	–16.15	0.00 (0.00–inf)	0.43*	1.54 (1.01–2.33)
log θ	0.57**			
log likelihood, d.f.	–839.48 (17)			

CI = Confidence interval; OR = odds ratio; d.f. = degrees of freedom. * p < 0.05, ** p < 0.01.

had a dmfs score larger than zero. A remedy to overcome this issue (using a bias-reduced logistic regression) is presented in, for example, Heinze and Schemper [2002] and Kosmidis and Firth [2009], and in the accompanying R code.

Since the zero counts and the nonzero counts are separated in the hurdle model, the relationship between a predictor variable and the amount of positive caries counts is relatively straightforward. For example, among Dutch girls who brush their teeth twice a day, have breakfast every day, take food and drinks less than 7 times a day, and have a Corah score below 13, the rate ratio of the mean caries counts from the untruncated (negative binomial) distributions of having a low-educated mother compared to a high(er)-educated mother is exp(0.30) = 1.36. The overall predicted caries count for this example child with a low-educated mother is 2.85, whereas the overall predicted caries count for a child with the same characteristics, except having a higher-educated mother, is 1.78. These positive predicted caries counts, as well as the predicted number of zeros and the predicted overall means, are easily computed using the predict() function in R.

We also inspected θ to confirm our choice between the Poisson hurdle and the negative binomial hurdle. As can be seen from table 3, θ indicated overdispersion (log θ = 0.51). Therefore, a negative binomial distribution shows a better fit, as was also indicated by the likelihood ratio test (table 2).

The interpretation of the coefficients from the zero-inflated model is more complex. The first column ('coefficient zero part') relates to the modelling of *excess zeros*, and the second column presents the probability of an individual with that particular characteristic being an excess zero. The modeling of the negative binomial is given in the third column, and the IRR is presented in the fourth column. The odds of observing an excess zero for someone who is not having breakfast every day, given the other variables in the model, is exp(–1.41) = 0.25 times the odds for someone who is having breakfast every morning. The interpretation of these 'excess' zeros is somewhat hard to grasp: these zeroes represent the log odds ratio of being in the 'not-at-risk-for-caries group'. Therefore, one cannot infer any conclusions from these parameters about observing no caries experience in the overall population. Following Preisser et al. [2012], the IRR severity for children with a low-educated mother compared to children with a higher-educated mother, when all other characteristics are coded as 0 (that is, being a girl, brushing their teeth twice a day, having breakfast every day, taking food and drinks less than 7 times a day, and having a Corah score below 13) is defined by

$$\frac{E\left(Y_i|x_{i1}=1, x_{i2}=0, x_{i3}=0, x_{i4}=0, x_{i5}=0, x_{i6}=0, x_{i7}=0\right)}{E\left(Y_i|x_{i1}=0, x_{i2}=0, x_{i3}=0, x_{i4}=0, x_{i5}=0, x_{i6}=0, x_{i7}=0\right)} = \exp\left(\beta_1\right)\frac{1+\exp\left(\gamma_0\right)}{1+\exp\left(\gamma_0+\gamma_1\right)} \qquad (8)$$

By filling in the estimates from table 4 into equation 8, the IRR for the overall effect of education level of the mother on caries severity is 1.59. One can easily use the predict() function in R to compute predicted means and means from the count and zero components.

Conclusion

Hurdle models and zero-inflated models are both regression techniques that could be used to model DMF data with a positively skewed distribution and a large number of zero values. Whereas zero-inflated models are becoming more and more adopted in caries research, the use of hurdle models clearly lags behind. With this paper, we have tried to bring hurdle models to the notice of caries researchers. Although hurdle models and zero-inflated models often show similar results, the interpretation of the hurdle model is easier. Because all zero values are modeled separately from the caries counts, inferences for the prevalence of caries incidence in the population can directly be made with a hurdle model. In zero-inflated models, on the other hand, zero values come from a mixture of two sources, and therefore estimates for the zero counts cannot be directly generalized to the overall population. Due to its relative simplicity compared to zero-inflated models, we believe that the use of hurdle models is of added value for caries researchers. We hope that this paper, along with the example data and R code, helps caries researchers to easily adopt hurdle models into their research.

Acknowledgment

The study 'Oral health in children and adolescents in the Netherlands' was financed by the Health Care Insurance Board (Zorginstituut Nederland), Diemen, The Netherlands. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

The data were analyzed by H.H., E.D., and A.Z. The paper was written by H.H., E.D., A.Z., and A.S.

Disclosure Statement

The authors have no competing interests to declare.

References

Akaike H: Information theory and an extension of the maximum likelihood principle; in Petrov BN, Csaki BF (eds): Second International Symposium on Information Theory. Budapest, Academiai Kiado, 1973, pp 267–281.

Böhning D, Dietz E, Schlattmann P: The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. J R Stat Soc Ser A 1999;162:195–209.

Campus G, Cagetti MG, Senna A, Blasi G, Mascolo A, Demarchi P, Strohmenger L: Does smoking increase risk for caries? A cross-sectional study in an Italian military academy. Caries Res 2011;45:40–46.

Cheung YB: Zero-inflated models for regression analysis of count data: a study of growth and development. Stat Med 2002;21:1461–1469.

Dusseldorp E, Kamphuis M, Schuller A: Impact of lifestyle factors on caries experience in three different age groups: 9, 15, and 21-year-olds. Community Dent Oral Epidemiol 2015; 43:9–16.

Heinze G, Schemper M: A solution to the problem of separation in logistic regression. Stat Med 2002;21:2409–2419.

Hu M, Pavlicova M, Nunes EV: Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. Am J Drug Alcohol Abuse 2011; 37:367–375.

Jahani Y, Eshraghian MR, Foroushani A, Nourijelyani K, Mohammad K, Shahravan A, Alam M: Effect of socio-demographic status on dental caries in pupils by using a multilevel hurdle model. Health 2013;5:1110–1116.

Javali S, Pandit P: Using zero-inflated models to analyze dental caries with many zeros. Indian J Dent Res 2010;21:480–485.

Kleiber C, Zeileis A: Visualizing Count Data Regressions Using Rootograms. Working Paper 2014–20. Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics. Innsbruck, Universität Innsbruck, 2014.

Kosmidis I, Firth D: Bias reduction in exponential family nonlinear models. Biometrika 2009;96: 793–804.

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

Lewsey JD, Thomson WM: The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. Community Dent Oral Epidemiol 2004;32:183–189.

Liu L, Zhang Y, Wu W, Cheng M, Li Y, Cheng R: Prevalence and correlates of dental caries in an elderly population in northeast China. PLoS One 2013;8:e78723.

Min YM, Agresti A: Random effect model for repeated measures of zero-inflated count data. Stat Model 2005;5:1–19.

Molenberghs G, Verbeke G: Likelihood ratio, score, and Wald test in a constrained parameter space. Am Stat 2012;61:22–27.

Mullahy J: Specification and testing of some modified count data models. J Econom 1986;33: 341–365.

Nobile C, Fortunato L, Bianco A, Pileggi C, Pavia M: Pattern and severity of early childhood caries in Southern Italy: a preschool-based cross-sectional study. BMC Public Health 2014;14:206.

Preisser JS, Das K, Benecha H, Stamm JW: Logistic regression for dichotomized counts. Stat Methods Med Res 2014 DOI: 10.1177/ 0962280214536893.

Preisser JS, Stamm JW, Long DL, Kincade ME: Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries Res 2012;46:413–423.

R Core Team: R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing, 2013.

Schuller AA, Poorterman JHG, van Kempen CPF, Dusseldorp E, van Dommelen P, Verrips GHW: Kies voor tanden: een onderzoek naar mondgezondheid en preventief tandheelkundig gedrag van jeugdigen. Tussenmeting 2009, een vervolg op de reeks TJZ-onderzoeken. Leiden, TNO, 2011.

Schwarz G: Estimating the dimension of a model. Ann Stat 1978;6:461–464.

Solinas G, Campus G, Maida C, Sotgiu G, Cagetti MG, Lesaffre E, Castiglia P: What statistical method should be used to evaluate risk factors associated with dmfs index? Evidence from the National Pathfinder Survey of 4-year-old Italian children. Community Dent Oral Epidemiol 2009;37:539–546.

Vuong Q: Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica 1989;57:307–334.

Zeileis A, Kleiber C, Jackman S: Regression models for count data in R. J Stat Softw 2008;27: 1–25.