ORIGINAL PAPER

Transition models for count data: a flexible alternative to fixed distribution models

Moritz Berger[1] · Gerhard Tutz[2]

Abstract

A flexible semiparametric class of models is introduced that offers an alternative to classical regression models for count data as the Poisson and Negative Binomial model, as well as to more general models accounting for excess zeros that are also based on fixed distributional assumptions. The model allows that the data itself determine the distribution of the response variable, but, in its basic form, uses a parametric term that specifies the effect of explanatory variables. In addition, an extended version is considered, in which the effects of covariates are specified nonparametrically. The proposed model and traditional models are compared in simulations and by utilizing several real data applications from the area of health and social science.

Keywords Count data · Smoothing · Transition model · Varying coefficients · Zero-inflated model

1 Introduction

In many applications the response variable of interest is a nonnegative integer or count which one wants to relate to a set of covariates. Classical regression models are the Poisson model and the Negative Binomial model, which can be embedded into the framework of generalized linear models (McCullagh and Nelder 1989). More general models use the generalized Poisson distribution, the double Poisson distribution or the Conway–Maxwell–Poisson distribution (Consul 1998; Zou et al. 2013; Sellers and Shmueli 2010). Specific models designed to account for excess zeros are the hurdle model and the zero-inflated model. Concise overviews of

✉ Moritz Berger
   moritz.berger@imbie.uni-bonn.de

[1] Institut für Medizinische Biometrie, Informatik und Epidemiologie, Medizinische Fakultät, Universität Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

[2] Ludwig-Maximilians-Universität München, Akademiestrasse 1, 80799 Munich, Germany

modeling strategies were given by Kleiber and Zeileis (2008), Hilbe (2011), Cameron and Trivedi (2013) and Hilbe (2014). Specific models using distributions beyond the classical ones were considered, among others, by Joe and Zhu (2005), Gschoessl and Czado (2006), Nikoloulopoulos and Karlis (2008), Rigby et al. (2008) and Böhning and van der Heijden (2009). More recently, existing count regression approaches were compared, for example, by Hayat and Higgins (2014), Payne et al. (2017) and Maxwell et al. (2018).

Most of the established models assume that a fixed distribution holds for the response variable conditional on the values of covariates, and the mean (and possibly the variance parameter) is linked to a linear function of the covariates. Various methods have been proposed to estimate the parametric effect of covariates on the counts (Cameron and Trivedi 2013). Specifying a distribution of the counts, however, can be rather restrictive, and the validity of inference tools depends on the correct specification of the distribution.

To address this issue, we propose a class of models that does not require specifying a fixed distribution for the counts. Instead, the form of the distribution is determined by parameters that reflect the tendency to higher counts, which has the effect that the fitted distribution is fully data-driven. For the estimation of the parameters that determine the distribution penalized maximum likelihood estimation procedures are introduced. The proposed models automatically account for zero inflation, which typically calls for more complex models, see, for example, Mullahy (1986a), Lambert (1992), Loeys et al. (2012). Interpretation of the parameters is kept simple, since the conceptualization uses that counts typically result from a process, where the final count is the result of increasing numbers.

The effect of covariates on the response is modeled by a linear term. This enables an easy interpretation of the regression coefficients in terms of multiplicative increases or decreases of the counts. The proposed models are semiparametric in nature, because the distribution of the response is modeled in a flexible way adapted to the data while the effect of covariates is modeled parametrically. The models are also extended to allow for smoothly varying coefficients in a nonparametric fashion.

A main advantage of the proposed model class is that it can be embedded into the framework of binary regression. This implies that standard software for maximum likelihood estimation of generalized additive models (Wood 2006) can be used for model fitting.

The rest of the article is organized as follows: In Sect. 2 classical models for count data are briefly reviewed. In Sect. 3 the model class is introduced and penalized maximum likelihood estimation methods are described. A simulation study investigating the properties of the proposed model class is presented in Sect. 4. Sections 5 and 6 are devoted to applications from the area of health and social science. In Sect. 6 the model is also extended to allow for more flexible effects of covariates, which are not necessarily restricted to linear effects. Details on implementation and software are given in Sect. 7. Section 8 summarizes the main findings of the article.

2 Classical models for count data

Let $Y_i \in \{0, 1, 2, \ldots\}$ denote the response variable and $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip})$ a vector of covariates of an i.i.d. sample with n observations. In generalized linear models (GLMs) one specifies a distributional assumption for $Y_i|\boldsymbol{x}_i$ and a structural assumption that links the mean $\mu_i = \mathrm{E}(Y_i|\boldsymbol{x}_i)$ to the covariates. The structural assumption in GLMs has the form

$$\mu_i = h(\boldsymbol{x}_i^T \boldsymbol{\beta}) \quad \text{or} \quad g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a set of real-valued coefficients, g is a known link function and $h = g^{-1}$ denotes the response function, see, for example, McCullagh and Nelder (1989), Fahrmeir and Tutz (2001).

2.1 Poisson and negative binomial model

Popular models for count data are the Poisson model and the Negative Binomial model. The Poisson model assumes that $Y_i|\boldsymbol{x}_i$ follows a Poisson distribution $P(\mu_i)$, where the mean and the variance are given by μ_i, respectively. The most widely used model uses the canonical link function by specifying

$$\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) \quad \text{or} \quad \log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{1}$$

In model (1) the expressions $\exp(\beta_1), \ldots, \exp(\beta_p)$ have a simple interpretation in terms of multiplicative increases or decreases of μ_i. More generally one can assume that $Y_i|\boldsymbol{x}_i$ follows a Negative Binomial distribution $\mathrm{NB}(v, \mu_i)$ with probability density function (p.d.f.)

$$P(Y_i = y_i|\boldsymbol{x}_i) = \frac{\Gamma(Y_i + v)}{\Gamma(v)\Gamma(Y_i + 1)} \left(\frac{\mu_i}{\mu_i + v}\right)^{y_i} \left(\frac{v}{\mu_i + v}\right)^v, \; y_i \in \{0, 1, 2, \ldots\},$$

where $\Gamma(\cdot)$ denotes the gamma function. The parameter v is a shape parameter that allows more flexible modeling of the variance than the Poisson distribution. Mean and variance are given by

$$\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) \quad \text{and} \quad \mathrm{var}(Y_i|\boldsymbol{x}_i) = \mu_i + \mu_i^2/v. \tag{2}$$

While the mean is the same as for the simple Poisson model, the variance exceeds the Poisson variances by μ_i^2/v, which may be seen as a limiting case ($v \to \infty$). As v is not related to the mean, it is assumed to be constant for all observations. For alternative formulations of the Negative Binomial model, see, for example, Cameron and Trivedi (2013).

2.2 Zero-inflated model

In some applications one observes more zero counts than is consistent with the Poisson or Negative Binomial model; then data display overdispersion through excess zeros. This happens in cases in which the population consists of two

subpopulations, the non-responders who are "never at risk" with counts $Y_i = 0$ and the responders who are at risk with counts $Y_i \in \{0, 1, \ldots\}$, see, for example, Cameron and Trivedi (2005), Cameron and Trivedi (2013). Formally, the zero-inflated density function is a mixture of distributions. With C_i denoting the class indicator of subpopulations ($C_i = 1$ for responders and $C_i = 0$ for non-responders) one obtains the mixture distribution

$$P(Y_i = y_i | \boldsymbol{x}_i) = P(Y_i = y_i | C_i = 0)\,\pi_i + P(Y_i = y_i | \boldsymbol{x}_i, C_i = 1)\,(1 - \pi_i),$$

where $\pi_i = P(C_i = 0)$ are the mixing probabilities. Typically one assumes that a classical count data model, for example the Poisson model (1), holds for the responders, that is, one assumes $Y_i | \boldsymbol{x}_i, C_i = 1 \sim P(\mu_i)$, and a binary model, for example the logistic model, determines class membership. Then the link between responses and covariates is determined by the two predictors

$$\log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \quad \text{and} \quad \text{logit}(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\gamma}, \tag{3}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ is a second set of real-valued coefficients.

Instead of the Poisson model also the Negative Binomial model can be used to represent responders, see, for example, Greene (1994). Alternative models, in which the Poisson distribution is replaced by the generalized Poisson distribution have been considered by Famoye and Singh (2003), Gupta et al. (2004), Famoye and Singh (2006), Czado et al. (2007) and Min and Czado (2010). Identifiability of zero-inflated models was investigated by Li (2012). Estimation procedures for zero-inflated models are available in the R package **pscl** (Zeileis et al. 2008).

2.3 Hurdle model

An alternative model that is able to account for excess zeros is the hurdle model (Mullahy 1986b; Creel and Loomis 1990). The model specifies two processes that generate the zero counts and the positive counts. It combines a truncated-at-zero count model (left-truncated at $Y_i = 1$) which is employed for positive counts and a binary model or censored count model (right-censored at $Y_i = 1$) which determines whether the response is zero or positive, i.e., if the "hurdle is crossed". Formally, the hurdle model is determined by

$$P(Y_i = y_i | \boldsymbol{x}_i) = \begin{cases} f_1(0|\boldsymbol{x}_i) & \text{if } y_i = 0, \\ f_2(Y_i|\boldsymbol{x}_i)\dfrac{1 - f_1(0|\boldsymbol{x}_i)}{1 - f_2(0|\boldsymbol{x}_i)} & \text{if } y_i > 0, \end{cases}$$

where f_1 determines the binary decision between zero and a positive response. If the hurdle is crossed, the response is determined by the truncated count model with p.d.f.

$$P(Y_i = y_i | \boldsymbol{x}_i) = f_2(y_i|\boldsymbol{x}_i)/(1 - f_2(0|\boldsymbol{x}_i)), \qquad y_i = 1, 2, \ldots.$$

If $f_1 = f_2$, the model collapses to the so-called *parent process* f_2. The model is quite flexible, because it allows for both under- and overdispersion. For example, in the

hurdle Poisson model where both f_1 and f_2 correspond to Poisson distributions with means μ_1 and μ_2, the link between responses and covariates is determined by the two predictors

$$\mu_{i1} = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) \quad \text{and} \quad \mu_{i2} = \exp(\boldsymbol{x}_i^T \boldsymbol{\gamma}). \tag{4}$$

The Poisson and geometric hurdle model have been examined by Mullahy (1986b), Negative Binomial hurdle models have been considered by Pohlmeier and Ulrich (1995). Zero inflated and hurdle models have also been generalized to multiple inflated models to allow for count-inflation at multiple values, see Giles (2007), and, more recently, Bocci et al. (2020).

Estimation procedures for hurdle models are available in the R package **pscl** (Zeileis et al. 2008).

3 The transition model for count data

In particular the Poisson and the Negative Binomial model have a simple structure with a clearly defined link between the covariates and the mean. The models however are rather restrictive since they assume that the distribution of $Y_i|\boldsymbol{x}_i$ is known and fixed. A strict parametric form is assumed for the whole support $\{0, 1, 2, \ldots\}$ and typically just the dependence of the mean on the predictors is specified by the model.

3.1 The basic transition model

The approach proposed here is much more flexible and does not assume a fixed distribution for the response. It focuses on the modeling of the transition between counts. In its simplest form, it assumes for $Y_i \in \{0, 1, 2, \ldots\}$

$$P(Y_i > r | Y_i \geq r, \boldsymbol{x}_i) = F(\theta_r + \boldsymbol{x}_i^T \boldsymbol{\beta}), \quad r = 0, 1, \ldots, \tag{5}$$

where $F(\cdot)$ is a fixed distribution function. The parameters θ_r represent intercept coefficients and $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ is a vector of regression coefficients. One models the transition probability $\delta_{ir} = P(Y_i > r | Y_i \geq r, \boldsymbol{x}_i)$, which is the conditional probability that the number of counts is larger than r (i.e., transition to a higher value than r) given the number of counts is at least r. These probabilities are determined by a classical binary regression model. For example, if $F(\cdot)$ is the logistic distribution function, one uses the binary logit model.

In general, a distribution of count data Y_i can be characterized by the probabilities on the support $\pi_{i0}, \pi_{i1}, \ldots$, where $\pi_{ir} = P(Y_i = r)$, *or* the (conditional) transition probabilities $\delta_{i0}, \delta_{i1}, \ldots$ given by

$$\delta_{ir} = P(Y_i > r | Y_i \geq r) = \frac{1 - \pi_{i0} - \cdots - \pi_{ir}}{1 - \pi_{i0} - \cdots - \pi_{i,r-1}}.$$

The transition model (5) specifies the transition probabilities. If no covariates are present, any discrete distribution with support $\{0, 1, 2, \ldots\}$ can be represented by

the model, determined by the intercept parameters $\theta_0, \theta_1, \ldots$. In the presence of covariates the intercepts represent the basic distribution of the counts, which is modified by the values of the covariates. Thus, the functional form of the count distribution is not restricted. In particular the response may follow a Poisson distribution or a Negative Binomial distribution. In addition the model is able to account for specific phenomena like excess zeros.

The parameters in model (5) have an easy interpretation depending on the function $F(\cdot)$. If one chooses the logistic distribution function one obtains the conditional transition to a higher number of counts in the form

$$\log\left(\frac{P(Y_i > r | Y_i \geq r, \boldsymbol{x}_i)}{1 - P(Y_i > r | Y_i \geq r, \boldsymbol{x}_i)}\right) = \theta_r + \boldsymbol{x}_i^T \boldsymbol{\beta},$$

and the regression coefficients $\beta_1, \ldots, \beta_p$ have the usual interpretation as in the common binary logit model. An alternative form is the representation as continuation ratios

$$\log\left(\frac{P(Y_i > r | \boldsymbol{x}_i)}{P(Y_i = r | \boldsymbol{x}_i)}\right) = \theta_r + \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{6}$$

The ratio $P(Y_i > r | \boldsymbol{x}_i)/P(Y_i = r | \boldsymbol{x}_i)$ compares the probability that the number of counts is larger than r to the probability that the number of counts is equal to r.

The basic assumption of model (6) is that the effect of covariates is the same for any given number of counts r. This property can also be seen as a form of strict stochastic ordering. That means, if one considers two population that are characterized by the covariate values $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$, one obtains

$$\frac{P(Y > r | \boldsymbol{x})/P(Y = r | \boldsymbol{x})}{P(Y > r | \tilde{\boldsymbol{x}})/P(Y = r | \tilde{\boldsymbol{x}})} = \exp((\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \boldsymbol{\beta}). \tag{7}$$

Thus the comparison of populations in terms of the transition odds $P(Y > r | \boldsymbol{x})/P(Y = r | \boldsymbol{x})$ does not depend on the counts. If, for example, the odds in population $\boldsymbol{x}$ are twice the odds in population $\tilde{\boldsymbol{x}}$ this holds for all r.

The modeling of transitions as specified in model (5) was used in various contexts before. In ordinal regression transition modeling is known under the name sequential model; in the logistic version it is called continuation ratio model (Agresti 2002; Tutz 2012). Its properties as an ordinal regression model have been investigated in particular by Maxwell et al. (2018). It is also used in discrete survival analysis, where one parameterizes the discrete hazard function $\lambda(r | \boldsymbol{x}_i) = P(Y_i = r | Y_i \geq r, \boldsymbol{x}_i)$ instead of the conditional probability of transition (Tutz and Schmid 2016).

Modeling of transitions, however, seems not to have been used in the modeling of count data. The main difference to the use in ordinal regression and discrete hazard modeling is that, in contrast to these models, the number of categories is not restricted. In ordinal models one typically uses up to ten categories, which limits the number of parameters. In count data, however, there is no restriction on the number of intercepts (i.e. the possible number of counts). Particularly, in extended models, where also the regression coefficients vary across categories (to be considered in

Sect. 6), the main problem is that the number of parameters cannot be handled by simple maximum likelihood estimation. For count data and fixed predictor value, model (5) is a Markov chain model of order one, because the probability of transition depends only on the previously obtained category. It is related to categorical time series, which have been investigated by Kaufmann (1987), Fahrmeir and Kaufmann (1987), and Kedem and Fokianos (2002).

3.2 Illustration of flexibility of the model

Before giving details of the fitting procedure we demonstrate the flexibility of the proposed transition model by a small benchmark experiment that was based on 100 replications. We generated samples of size $n = 100$ with the outcome values drawn from (i) a Poisson distribution, $y_i \sim Po(\mu_i = 5)$, $i = 1, \ldots, n$, and (ii) a Negative Binomial distribution, $y_i \sim NB(v = 5/8, \mu_i = 5)$, $i = 1, \ldots, n$, which equals variance $\text{var}(y_i) = 45$.

Figure 1 shows the estimated probability density functions for 10 randomly chosen replications (upper and middle panel) and the average estimated probability density function over all 100 replications (lower panel) obtained from fitting the transition model and the true data-generating model (Poisson or Negative Binomial) to the data, respectively. In both cases, it is seen that the transition model is well able to capture the underlying distribution. In particular, the average estimated p.d.f. and the true p.d.f. (black line) closely coincide for both distributions (lower panel).

3.3 Maximum likelihood estimation

For i.i.d. observations $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, the log-likelihood has the simple form

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \log(\pi_{ir}),$$

where $\boldsymbol{\alpha}^T = (\theta_0, \theta_1, \ldots, \boldsymbol{\beta}^T)$ collects all parameters and $\pi_{ir} = P(Y_i = r | \boldsymbol{x}_i)$ is the probability of observing category r, which for the transition model (5) is given by

$$
\begin{aligned}
\pi_{ir} &= P(Y_i = r | Y_i \geq r, \boldsymbol{x}_i) \prod_{s=0}^{r-1} P(Y_i > s | Y_i \geq s, \boldsymbol{x}_i) \\
&= \left(1 - F(\theta_r + \boldsymbol{x}_i^T \boldsymbol{\beta})\right) \prod_{s=0}^{r-1} F(\theta_s + \boldsymbol{x}_i^T \boldsymbol{\beta}).
\end{aligned}
\tag{8}
$$

For the model considered here it is helpful to represent the data in a different way. One considers the underlying Markov chain $Y_{i0}, Y_{i1}, Y_{i2}, \ldots$, where $Y_{ir} = I(Y_i = r)$ with $I(\cdot)$ denoting the indicator function ($I(a) = 1$, if a holds, $I(a) = 0$ otherwise). Then the log-likelihood can be given in the form
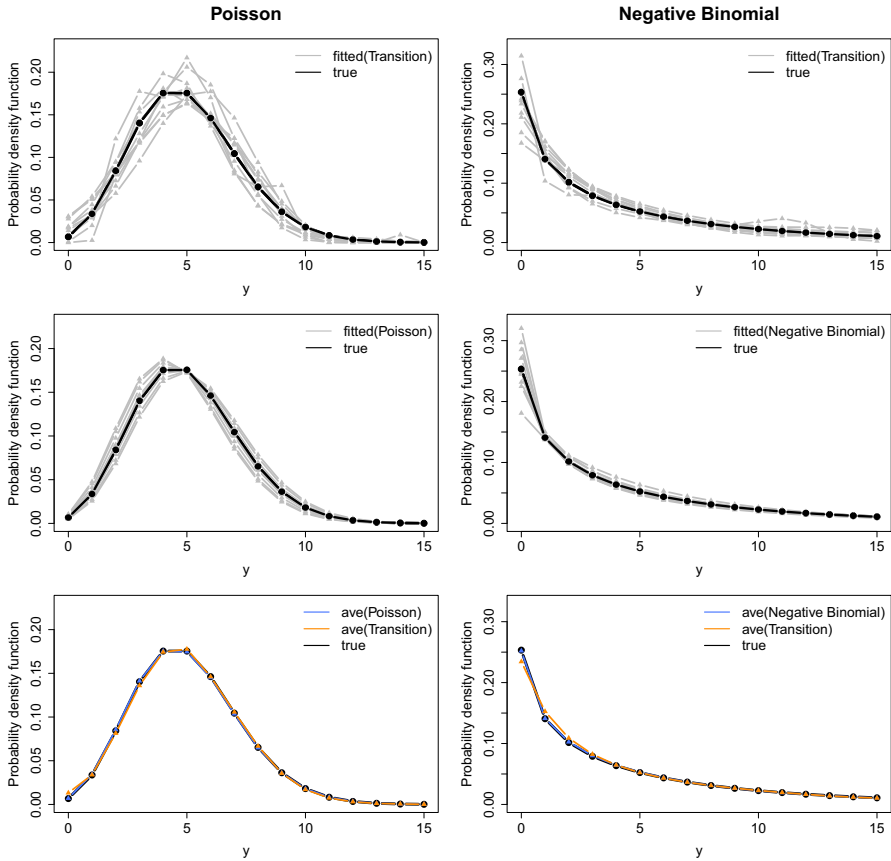
Fig. 1 Estimated probability density functions when fitting the transition model and the true data-generating model to samples drawn from a Poisson distribution (left) and samples drawn from a Negative Binomial distribution (right). The upper and middle panels show the estimates of 10 randomly chosen replications when the transition model (upper panel) and the true data generating model (middle panel) are fitted. The lower panel shows the respective average estimates obtained from 100 simulation runs. The black lines correspond to the true probability density function

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{s=0}^{Y_i} Y_{is} \log(\pi_{is}).$$

By using (8) it can be rewritten as

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{s=0}^{Y_i} Y_{is} \log\big(1 - F(\theta_s + \boldsymbol{x}_i^T \boldsymbol{\beta})\big) + (1 - Y_{is}) \log\big(F(\theta_s + \boldsymbol{x}_i^T \boldsymbol{\beta})\big). \qquad (9)$$

In (9) the realizations of the Markov chain $Y_{i0}, Y_{i1}, \ldots, Y_{iY_i}$ up to the observed response have the form $(Y_{i0}, Y_{i1}, \ldots, Y_{iY_i})^T = (0, 0, \ldots, 0, 1)$. They can be seen as dummy variables for the response or as binary variables that indicate if transition to the next category occurred or not. The value $Y_{ir} = 0$, $r < Y_i$, denotes that the

transition to a higher category than r occurred. If one wants to indicate transition as 0-1 variable with 1 denoting transition to a higher category one uses $\tilde{Y}_{ir} = 1 - Y_{ir}$ yielding the log-likelihood

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{s=0}^{Y_i} \tilde{Y}_{is} \log\big(F(\theta_s + \boldsymbol{x}_i^T \boldsymbol{\beta})\big) + (1 - \tilde{Y}_{is}) \log\big(1 - F(\theta_s + \boldsymbol{x}_i^T \boldsymbol{\beta})\big), \quad (10)$$

which is obviously equivalent to the log-likelihood of the binary response model $P(\tilde{Y}_{ir} = 1 | \boldsymbol{x}_i) = F(\theta_r + \boldsymbol{x}_i^T \boldsymbol{\beta})$ for observations $\tilde{Y}_{i0}, \tilde{Y}_{i1}, \ldots, \tilde{Y}_{iY_i}$. Thus the model can be fitted by using maximum likelihood methods for binary data that encode the sequence of transitions up to the observed response.

3.3.1 Penalized maximum likelihood estimation

Maximum likelihood (ML) estimators tend to fail because model (5) contains many parameters, in particular the number of intercepts becomes large unless the counts are restricted to very small numbers. Therefore alternative estimators are needed. We will use penalized maximum likelihood estimates. Then instead of the log-likelihood (10) one maximizes the penalized log-likelihood

$$\ell_p(\boldsymbol{\alpha}) = \ell(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}), \quad (11)$$

where $\ell(\cdot)$ is the common log-likelihood of the model and $J(\boldsymbol{\alpha})$ is a penalty term that penalizes specific structures in the parameter vector. The parameter λ is a tuning parameter that specifies how serious the penalty term has to be taken. Since the intercept parameters θ_r determine the dimensionality of the estimation problem the penalty is used to regularize these parameters.

A reasonable assumption on the θ-parameters is that they are changing slowly over categories. A penalty that enforces smoothing over response categories uses the squared differences between adjacent categories, Let $M_{\max}$ denote the maximal value that has been observed, that is, $M_{\max} = \max\{Y_i\}$. Then one uses the penalty

$$J(\boldsymbol{\alpha}) = \sum_{s=0}^{M} (\theta_s - \theta_{s-1})^2, \quad (12)$$

where M is larger than $M_{\max}$, for example, M can be chosen as the integer closest to $1.2\,M_{\max}$. When maximizing the penalized log-likelihood one obtains $\theta_{M_{\max}} = \theta_{M_{\max}+1} = \ldots = \theta_M$. It is important to choose M larger than $M_{\max}$ to account for possibly larger future observations and to avoid irregularities at the boundaries. If one uses $\lambda = 0$ in (11) one obtains the ML estimates. In the extreme case $\lambda \to \infty$ all parameters obtain the same value.

An alternative, more general smoothing technique uses penalized splines as proposed by Eilers and Marx (1996). Let the θ-parameters be specified as a smooth function over categories by using an expansion in basis functions of the form

$$\theta_r = \sum_{k=1}^{m} \gamma_k \phi_k(r) \,,$$

where $\phi_k(\cdot)$ are fixed basis functions, and m denotes the number of basis functions. We will use B-splines (Eilers and Marx 1996) on equally spaced knots in the range $[0, M]$, where M is again a larger value than the maximal observed response. The penalty now does not refer to the θ-parameters themselves but to the γ-parameters. A flexible form is

$$J(\boldsymbol{\alpha}) = \sum_{k=d+1}^{m} (\Delta^d \gamma_k)^2 \,, \tag{13}$$

where Δ^d is the difference operator, operating on adjacent B-spline coefficients, that is, $\Delta \gamma_k = \gamma_k - \gamma_{k-1}, \Delta^2 \gamma_k = \Delta(\gamma_k - \gamma_{k-1}) = \gamma_k - 2\gamma_{k-1} + \gamma_{k-2}$. The method is referred to as *P-splines* (for penalized splines). P-splines are strong tools that have various advantages. They are very flexible and allow for different polynomial degrees of the basis functions and difference operators (Δ^d). Since the basis functions are strictly local single fitted basis functions have no effect on remote areas. Moreover, one obtains simple polynomials if the smoothing parameter takes large values.

3.3.2 Embedding into the framework of varying-coefficients models

The transition variables $(\tilde{Y}_{i0}, \tilde{Y}_{i1}, \ldots, \tilde{Y}_{iY_i})^T = (1, 1, \ldots, 1, 0)$ in (10) are binary variables. The log-likelihood is the same as for binary response models of the form $P(\tilde{Y}_{ir} = 1|\boldsymbol{x}_i) = F(\theta_r + \boldsymbol{x}_i^T \boldsymbol{\beta})$. The binary models can also be seen as varying-coefficients models of the form $P(\tilde{Y}_{ir} = 1|\boldsymbol{x}_i) = F(\beta(r) + \boldsymbol{x}_i^T \boldsymbol{\beta})$, where $\beta(r) = \theta_r$ is an unknown function of the counts, that is, the intercepts vary across counts. By considering the counts as an explanatory variable one may treat the models as specific varying-coefficients models and use existing software for fitting (see Sect. 7 for further details).

3.4 Selection of smoothing parameter and prediction accuracy

For the choice of the tuning parameter (for example by resampling or cross-validation) a criterion for the accuracy of prediction is needed. A classical approach in linear models is to estimate the mean and compare it to the actually observed response by using the quadratic distance. But since the whole distribution given a fixed covariate is estimated it is more appropriate to compare the estimated distribution to the degenerated distribution that represents the actual observation by using loss functions.

Candidates are the quadratic loss $L_2(\boldsymbol{\pi}_i, \hat{\boldsymbol{\pi}}_i) = \sum_r (\pi_{ir} - \hat{\pi}_{ir})^2$ and the Kullback–Leibler loss $L_{KL}(\boldsymbol{\pi}_i, \hat{\boldsymbol{\pi}}_i) = \sum_r \pi_{ir} \log(\pi_{ir}/\hat{\pi}_{ir})$, where the vectors $\boldsymbol{\pi}_i = (\pi_{i0}, \pi_{i1}, \ldots)^\top$ and $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i0}, \hat{\pi}_{i1}, \ldots)^\top$ denote the true and estimated probabilities. When the true

probability vector is replaced by the 0-1 vector of observations $Y_i = (Y_{i0}, Y_{i1}, Y_{i2}, \ldots)^\top$ one obtains for the quadratic loss the Brier score

$$L_2(Y_i, \hat{\pi}_i) = (1 - \hat{\pi}_{iY_i})^2 + \sum_{r \neq Y_i} \hat{\pi}_{ir}^2.$$

For the Kullback–Leibler loss one obtains the logarithmic score $L_{KL}(Y_i, \hat{\pi}_i) = -\log(\hat{\pi}_{i,Y_i})$. The latter has the disadvantage that the predictive distribution $\hat{\pi}_i$ is only evaluated at the observation Y_i. Therefore, it does not take the whole predictive distribution into account. As Gneiting and Raftery (2007) postulate, a desirable predictive distribution should be as sharp as possible and well calibrated. Sharpness refers to the concentration of the distribution and calibration to the agreement between the distribution and the observations. For count data, a more appropriate loss function derived from the continuous ranked probability score (Gneiting and Raftery 2007), which will also be used in the simulation and the applications, is

$$L_{RPS}(Y_i, \hat{\pi}_i) = \sum_r (\hat{\pi}_i(r) - I(Y_i \leq r))^2, \tag{14}$$

where $\hat{\pi}_i(r) = \hat{\pi}_{i0} + \hat{\pi}_{i1} + \cdots + \hat{\pi}_{ir}$ represents the cumulative distribution. It was also used by Czado et al. (2009) for the predictive assessment of count data.

4 Simulation study

To further demonstrate the flexibility and to show the added value of the proposed transition model compared to classical models we extended the numerical experiment shown in Sect. 3.2.

We generated data of size $n = 100$ from (a) a Poisson model with mean $\mu_i = 5$, $i = 1, \ldots, n$, (b) a Negative Binomial model with parameters $\mu_i = 5$, $v = 5/8$, (c) a zero-inflated Poisson model with parameters $\mu_i = 6.25$ and $\pi_i = 0.2$, (d) a zero-inflated Negative Binomial model with parameters $\mu_i = 6.25$, $v = 7/8$ and $\pi_i = 0.2$, and (e) a transition model with $M_{\max} = 21$ and decreasing intercepts $\theta_0 = -2$, $\theta_1 = -2.14$, …, $\theta_6 = -2.86$, $\theta_7 = -3$, and increasing intercepts $\theta_8 = -1$, $\theta_9 = -0.92$, …, $\theta_{19} = -0.08$, $\theta_{20} = 0$. The latter scenario results in data, where the distribution of the outcome values has two spikes. Figure 2 shows the distribution of one example data set.

In each of the five settings we generated a learning and test data set (100 replications), fitted all models to the learning data and evaluated the performance by calculating the ranked probability score (14) on the test data.

The results (given in Fig. 3) nicely illustrate that the Poisson distribution is quite rigid compared to the Negative Binomial distribution, as the performance of the Poisson model and the zero-inflated Poisson model deteriorates in scenarios (b) and (d). It is remarkable that the transition model performs consistently well and equal to the true data-generating model in the four scenarios (a) to (d). This underlines the high flexibility of our proposed method. In scenario (e) the
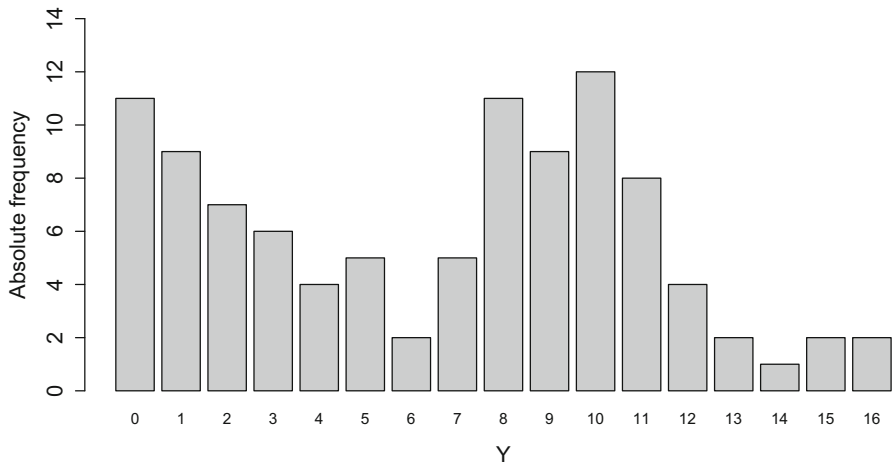
Fig. 2 Distribution of the outcome values for one exemplary data set in scenario (e) of the simulation study

transition model outperforms all classical parametric models (even the zero-inflated Negative Binomial model), which are less well able to capture the structure of the data. Our semiparametric approach therefore appears advantageous in such a more complex scenario.

5 Applications

To illustrate the usefulness of the proposed transition model, we present the results of three data examples comparing the various approaches introduced in the previous sections. Specifically, we consider

(i)   the Poisson model, hereinafter referred to as *Poisson*,

(ii)   the Negative Binomial model, referred to as *NegBin*,

(iii)   the zero-inflated model (3) using a Poisson model for the responders and a logit model to determine the class membership, referred to as *Zero (Poisson)*,

(iv)   the zero-inflated model (3) using a Negative Binomial model for the responders and a logit model to determine the class membership, referred to as *Zero (NegBin)*

(v)   the hurdle model (4) using a logit model for f_1 and a Poisson model for f_2, referred to as *Hurdle (Poisson)*,

(vi)   the hurdle model (4) using a logit model for f_1 and a Negative Binomial model for f_2, referred to as *Hurdle (NegBin)*,

(vii)   the transition model with a quadratic difference penalty (12) on the intercepts, referred to as *QuadPen* and

(a) Poisson

(b) NegBin

(c) Zero (Poisson)

(d) Zero (NegBin)
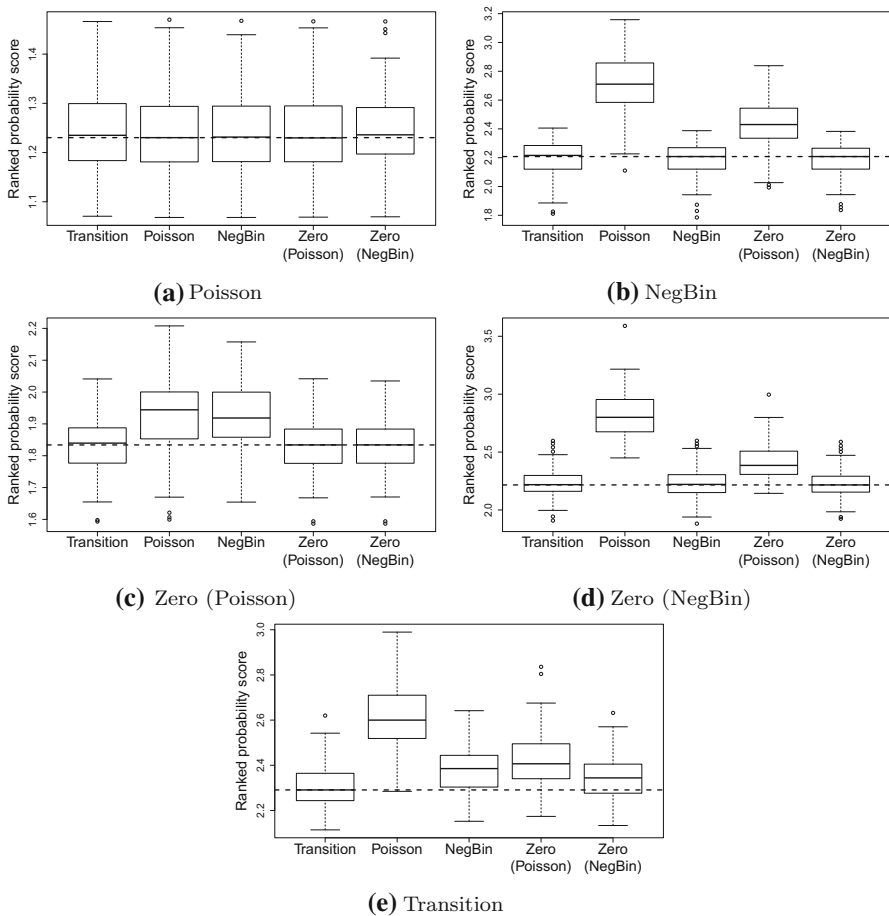
(e) Transition

Fig. 3 Results of the simulation study. The boxplots show the ranked probability score (for $r \in \{0, \ldots, 10\}$) of the transition model (left) and the four classical models (right). The median value of the true data generating model (indicated in the subheadings) is marked by a dashed line, respectively

    (viii)   the transition model, where the intercepts are expanded in cubic B-splines using a first order difference penalty (13), referred to as *P-Splines*.

To determine the optimal smoothing parameters λ for models (vii) and (viii) and to compare the predictive performance of all eight approaches we used the ranked probability score (14). For this purpose, we repeatedly (100 replications) generated subsamples without replacement containing 2/3 of the observations in the original data and computed the ranked probability score from the remaining test data sets (i.e., from 1/3 of the original data).

5.1 Absenteeism from school

We first consider a sociological study on children in Australia. The data set is available in the R package **MASS** (Venables and Ripley 2002) and was initially analysed by Aitkin (1979). The data consists of a sample of 146 children from New South Wales, Australia. The outcome of interest is the number of days a child was absent from school in one particular school year ($M_{max} = 81$). The unconditional distribution of the outcome for values $Y_i \in \{0, \ldots, 50\}$ is shown in Fig. 4 (eight observations had counts > 50). The covariates included in the models are Aboriginal ethnicity (Eth; 0: no, 1: yes), gender (Sex; 0: female, 1: male), the educational stage (Edu; 0: primary, 1: first form, 2: second form, 3: third form) and a learner status (Lrn; 0: average, 1: slow).

Figure 5 shows the results of the resampling experiment comparing the eight different approaches. In the case of the zero-inflated and Hurdle models the covariates were only included in the predictor for the responders/positive counts, respectively.

The ranked probability score was computed from the test data sets (containing 46 observations each) for outcome values $r \in \{0, \ldots, 30\}$. This range of values was chosen to ensure outcome values up to $r = 30$ in each of the learning and test samples. When fitting the transition models the minimal ranked probability score (averaged over 100 replications) was obtained for $\lambda = 23$ (QuadPen) and $\lambda = 466$ (P-Splines), respectively. It is seen that the Negative Binomial models (middle panel) and the two transition models (right panel) outperform the Poisson models with only minor differences between them. This result indicates that the Negative Binomial distribution is much more appropriate than the Poisson distribution for the analysis of this data set. Also accounting for excess zeros in the Negative Binomial model does not show an additional benefit. Importantly, the performance of the flexible transition model in terms of accuracy of prediction is the same as for the Negative Binomial model. Like in the simulations (b) and (d), the transition model is equally able to capture the essential characteristics of the data.
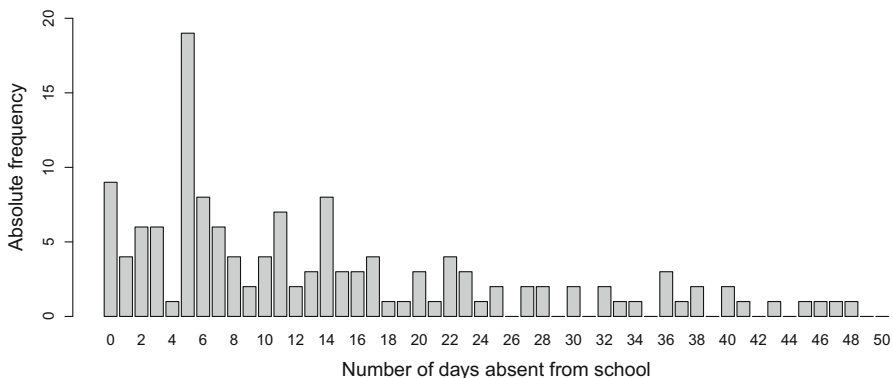

Fig. 4 Distribution of the outcome variable in the absenteeism from school data ($n = 146$)
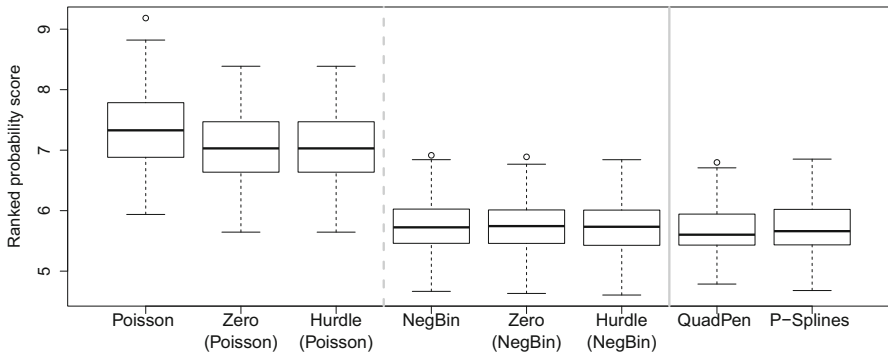
Fig. 5 Analysis of the absenteeism from school data. The boxplots show the ranked probability score (for $r \in \{0, \ldots, 30\}$) of the six classical models (left and middle panel) and the two transition models (right panel). All methods were fitted to 100 subsamples without replacement of size 100 and evaluated on the remaining 46 observations each

The estimated coefficients $\hat{\boldsymbol{\beta}}$ and the corresponding standard errors and z-values obtained by the Negative Binomial model and the transition model with penalized B-splines using the whole sample of $n = 146$ children are given in Table 1. Overall, the results of both models widely coincide. From the z-values it can be derived that only ethnicity has a significant effect on the outcome (at the 5% type 1 error level). Based on the Negative Binomial model, the expected number of days a child is absent from school is increased by the factor $\exp(0.569) = 1.766$ for the group of aboriginal children. In terms of the transition model, the continuation ratio (defined in (7)) is increased by the factor $\exp(0.585) = 1.795$, indicating higher counts in the group of aboriginal children.

5.2 Demand for medical care

Deb and Trivedi (1997) analyzed the demand for medical care for individuals, aged 66 and over, based on a dataset from the U.S. National Medical Expenditure survey

Table 1 Analysis of the absenteeism from school data

	NegBin			P-Splines		
	Coef	SE	z-value	Coef	SE	z-value
Eth	0.569	0.153	3.713	0.585	0.178	3.291
Sex	0.082	0.160	0.515	0.082	0.185	0.446
Edu:1	− 0.448	0.240	− 1.870	− 0.470	0.266	− 1.764
Edu:2	0.088	0.236	0.373	0.087	0.271	0.321
Edu:3	0.357	0.248	1.437	0.368	0.277	1.329
Lrn	0.292	0.186	1.566	0.309	0.205	1.507

Parameter estimates, standard errors and z-values obtained from fitting the Negative Binomial model (left) and the transition model with penalized B-splines (right)

in 1987/88. The data ("NMES1988") are available from the R package **AER** (Kleiber and Zeileis 2008). Like Zeileis et al. (2008) we consider the number of physician/non-physician office and hospital outpatient visits (Ofp) as outcome variable. The covariates used in the present analysis are the self-perceived health status (Health; 0: poor, 1: excellent), the number of hospital stays (Hosp), the number of chronic conditions (Numchron), age, marital status (Married; 0: no, 1: yes), and number of years of education (School). Since the effects vary across gender, we restrict consideration to male patients ($n = 356$). Figure 6 shows the unconditional distribution of the outcome $Y_i \in \{0, \ldots, 40\}$.

The ranked probabilty scores for outcome values $r \in \{0, \ldots, 30\}$ obtained from the approaches (i) to (viii) are shown in Fig. 7. Again, in the case of the zero-inflated and Hurdle models the covariates were only included in the predictor for the responders/positive counts, respectively. For the transition models the optimal smoothing parameters were $\lambda = 5$ (QuadPen) and $\lambda = 16$ (P-Splines). Similar to the previous example the Negative Binomial model (fourth boxplot) and the two transition models (seventh and eighth boxplot) performed considerably better than the Poisson model and at least as good as the models accounting for excess zeros.

The estimated coefficients $\hat{\boldsymbol{\beta}}$ and the corresponding standard errors and z-values obtained by the Negative Binomial model and the transition model with penalized B-splines using the whole sample of $n = 356$ patients are given in Table 2. Again, the two models yielded very similar results. An excellent health status reduced the expected number of visits, whereas the number of hospital stays and the number of years of education significantly increased the expected number of visits. Figure 8 shows the fitted smooth function of the θ-parameters obtained by the transition model with penalized B-splines, which represents the basic distribution of the counts. The function reveals decreasing coefficients with a local peak at ~ 20 visits.
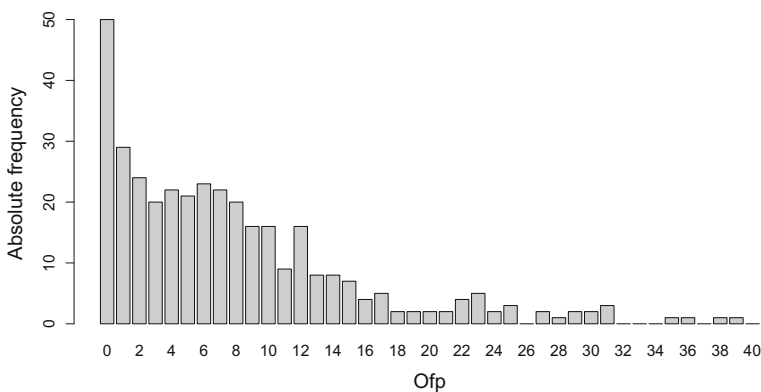

Fig. 6 Distribution of the outcome variable Ofp measured in the National Medical Expenditure survey ($n = 356$)
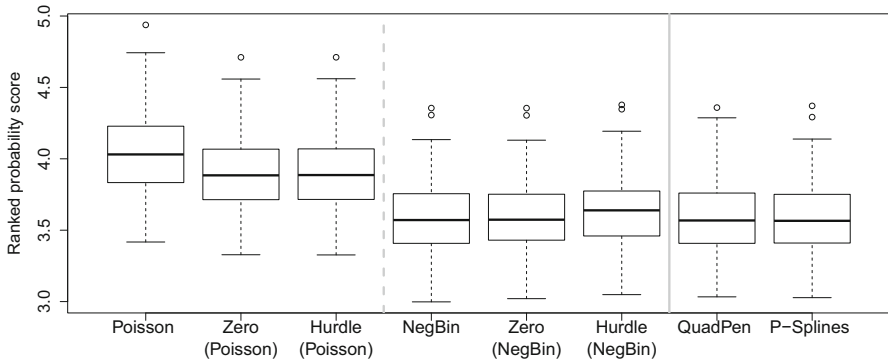
Fig. 7 Analysis of the medical care data. The boxplots show the ranked probability score (for $r \in \{0, \ldots, 30\}$) of the six classical models (left and middle panel) and the two transition models (right panel). All methods were fitted to 100 subsamples without replacement of size 237 and evaluated on the remaining 119 observations each

Table 2 Analysis of the medical care data

	NegBin			P-Splines		
	Coef	SE	z-value	Coef	SE	z-value
Health	− 0.681	0.140	− 4.863	− 0.794	0.158	− 5.021
Hosp	0.164	0.053	3.092	0.197	0.068	2.878
Numchron	0.058	0.040	1.452	0.057	0.043	1.312
Age	0.024	0.080	0.300	0.031	0.094	0.325
Married	0.074	0.122	0.608	0.067	0.138	0.482
School	0.042	0.013	3.259	0.045	0.014	3.127

Parameter estimates, standard errors and z-values obtained from fitting the Negative Binomial model (left) and the transition model with penalized B-splines (right)

Fig. 8 Analysis of the medical care data. Smooth function of the θ-parameters obtained from fitting the transition model with penalized B-splines ($\lambda = 16$)
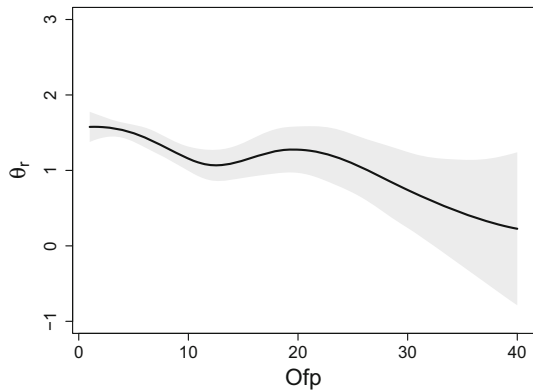
6 Transition model with varying coefficents

An extended form of the transition model assumes for $Y_i \in \{0, 1, 2, \ldots\}$

$$P(Y_i > r|Y_i \geq r, \boldsymbol{x}_i) = F(\theta_r + \boldsymbol{x}_i^T \boldsymbol{\beta}_r), r = 0, 1, \ldots, \tag{15}$$

where the regression coefficients $\boldsymbol{\beta}_r^T = (\boldsymbol{\beta}_{1r}, \ldots, \boldsymbol{\beta}_{pr})$ may vary over categories, i.e. number of counts. The parameter $\boldsymbol{\beta}_{jr}$ represents the weight on variable j for the transition to higher categories than r. The model does not assume that odds are proportional, and the stochastic ordering property (7) no longer holds. This makes the model much more flexible but strongly increases the number of parameters. A reduced number of effective parameters is obtained by assuming that they can be represented by basis functions, which implies that the parameters vary slowly across categories. Then the β-parameters are represented by

$$\boldsymbol{\beta}_{jr} = \sum_{k=1}^{m} \gamma_{jk} \phi_k(r), \quad j = 1, \ldots, p,$$

where $\phi_k(r)$ are fixed basis functions. The whole predictor of the model becomes

$$\log\left(\frac{P(Y_i > r|\boldsymbol{x}_i)}{P(Y_i = r|\boldsymbol{x}_i)}\right) = \sum_{k=1}^{m} \gamma_k \phi_k(r) + \sum_{j=1}^{p} \sum_{k=1}^{m} x_{ij} \gamma_{jk} \phi_k(r).$$

Estimation is again based on penalized maximum likelihood approaches with a penalty term on the differences of the γ-parameters on basis functions. Then, individual tuning parameters λ_j are used for each covariate to weight the sum of differences of the γ-parameters, respectively.

Extended model for excess zeros

A specific model with varying coefficients that is tailored to the case of excess zeros is obtained if one separates the first transition from all the other transitions. In the model

$$P(Y_i > 0|\boldsymbol{x}_i) = F(\theta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_0),$$
$$P(Y_i > r|Y_i \geq r, \boldsymbol{x}_i) = F(\theta_r + \boldsymbol{x}_i^T \boldsymbol{\beta}), \quad r = 1, 2, \ldots, \tag{16}$$

the first transition is determined by the parameter vector $\boldsymbol{\beta}_0$ while the other transitions are determined by the parameter vector $\boldsymbol{\beta}$. In model (16) one has varying coefficients $\beta(r) = \theta_r$ in the second equation of the model that vary across categories $r = 1, 2, \ldots$. These can again be fitted using a quadratic difference penalty or penalized B-splines as described for the basic model in Sect. 3. The model is a special case of model (15). It postulates a rather simple structure of varying coefficients by distinguishing just between the first transition and all other transitions. As in zero-inflated count models and hurdle models one specifies separate effects that model the occurrence of excess zeros. The model is referred to as *P-Splines (Zero)* in the following. It is useful if zero-inflation is suspected but a comparatively simple transition model with fixed parameter values can be used for all transitions beyond zero.

6.1 Absenteeism from school (contd.)

Let us again consider the study on schoolchildren in Australia. Figure 9 shows the results when fitting the extended transition model (15) in its most general form, that means the coefficients of all covariates are expanded in cubic B-splines using a first order difference penalty.

As in the basic transition model (cf. Table 1) the estimated functions indicate non-significant constant effects for girls compared to boys and for second form pupils compared to primary pupils (Edu:2). However, there are significant non-linear effects for ethnicity (which was also significant in the basic model) as well as for first form pupils (Edu:1) and learner status.

As an example, let us consider how slow learners compare to average learners (upper right panel of Fig. 9). It is seen that the continuation ratio increases by the factor $\exp(0.024) = 1.024$ $(r = 0)$ up to the factor $\exp(0.726) = 2.067$ $(r = 23)$, which doubles the probability for a higher count. That means the type of learner has a stronger impact on the days of absence if the student already has been absent for many days.

6.2 Demand for medical care (contd.)

The results from fitting the extended model (16) accounting for excess zeros to the data of the National Medical Expenditure survey are shown in Table 3 and Fig. 10. The first transition is referred to as *Zero* while the other transitions are referred to as the *Non-Zero* art of the model. There are remarkable differences compared to the
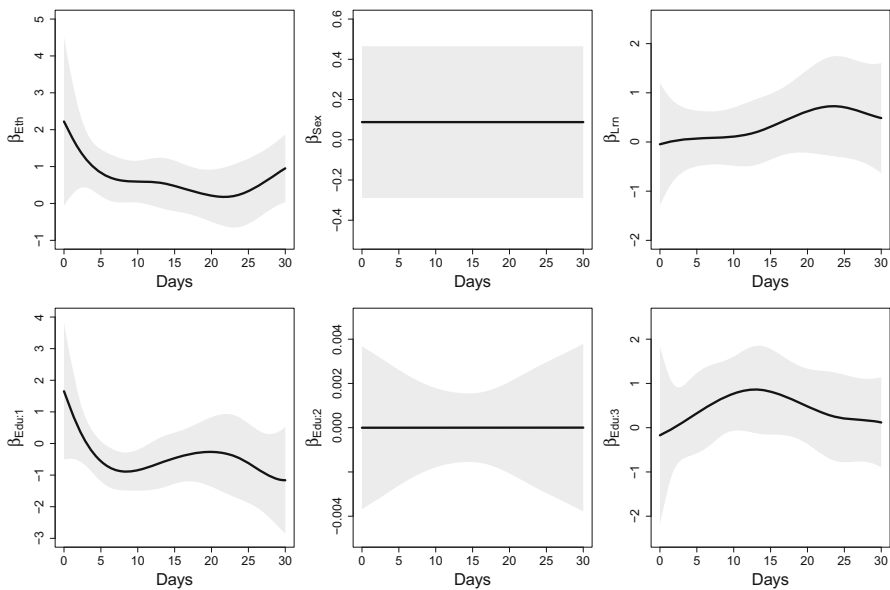

Fig. 9 Analysis of the absenteeism from school data. Smooth estimates β_r (for $r \in \{0, \ldots, 30\}$) obtained from fitting the extended transition model with varying coefficients in all covariates

Table 3 Analysis of the medical care data

	Zero			Non-Zero		
	Coef	SE	z-value	Coef	SE	z-value
θ_0	− 6.421	2.314	− 2.774	–	–	–
Health	− 0.375	0.466	− 0.806	− 0.749	0.171	− 4.379
Hosp	0.608	0.373	1.631	0.178	0.069	2.574
Numchron	0.591	0.187	3.161	0.015	0.045	0.336
Age	0.394	0.270	1.456	− 0.045	0.102	− 0.440
Married	0.512	0.361	1.417	− 0.007	0.152	− 0.044
School	0.161	0.041	3.918	0.025	0.016	1.591

Parameter estimates, standard errors and z-values obtained from fitting the extended transition with penalized B-splines accounting for excess zeros

Fig. 10 Analysis of the medical care data. Smooth function of the θ-parameters obtained from fitting the extended transition model with penalized B-splines ($\lambda = 16$)
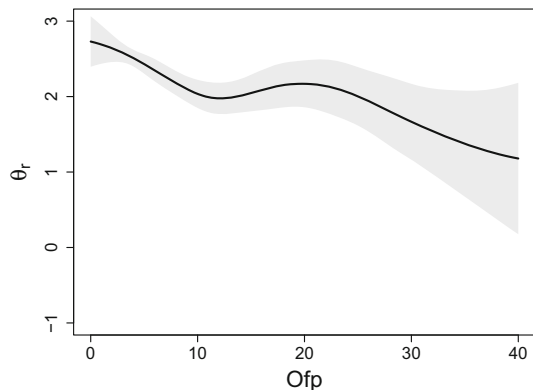
previous results in Table 2: (i) the number of years of education (school) had a significant effect on the first transition, but there was no evidence for an effect on the other transitions (z-value = 1.591), (ii) the number of chronic conditions, which was not significant in the basic model, showed a significant positive effect ($\hat{\beta}_0 = 0.591$) on the first transition (driving the decision to consult a doctor or not), and (iii) an excellent health status and the number of hospital stays had significant effects only in the part of the model that models the transition to higher categories given the number of visits was already above zero (Non-Zero). Figure 10 shows how the θ-parameters decrease over categories.

The ranked probability score of the extended model (evaluated on the test data sets and averaged over 100 replications) was 3.562, which indicates a slightly better predictive performance then all the previously considered models (cf. Fig. 7).

6.3 Boating trips

As third example we consider data based on a survey in 1980 administered to $n = 659$ leisure boat owners in eastern Texas, which is available from R package **AER** (Kleiber and Zeileis 2008) and was analyzed before by Ozuna and Gomez (1995). Here, the outcome of interest is the number of recreational boating trips to Lake Somerville, $Y_i \in \{0, \ldots, 40\}$. Note that 417 (63%) observations take the value zero, which calls for a model accounting for excess zeros (two extreme observations with outcome values 50 and 88 were excluded, as they seem implausible in the view of 52 weaks per year).

The five covariates used in the present analysis are the facility's subjective quality ranking (Quality; 1: very negative—5: very positive), an indicator, if the individual did water-skiing at the lake (Ski; 0: no, 1:yes), the annual household income (Income; in 1000 USD), an indicator, if the individual payed an annual user fee at the lake (Userfee; 0: no, 1: yes) and the expenditure when visiting the lake (Cost; in USD).

Next to the approaches (i)–(viii), we also considered the extended transition P-Splines (Zero) model (16). Here we included the covariates in both predictors when fitting zero-inflated and Hurdle models, which is in the same spirit as for the extended transition model, where zero and non-zero counts are separately determined by the covariates. Note that, due to a quasi-complete separation of the outcome with regard to Userfee (all individuals paying a user fee had > 0 counts), we excluded Userfee from the zero part of the models, respectively. Because of computational problems in many subsamples causing non-convergence, we had to exclude the zero-inflated Negative Binomial model from the analysis. Figure 11 shows the ranked probability scores computed from the test data sets (containing 219 observations each) for outcome values $r \in \{0, \ldots, 30\}$. It is seen that
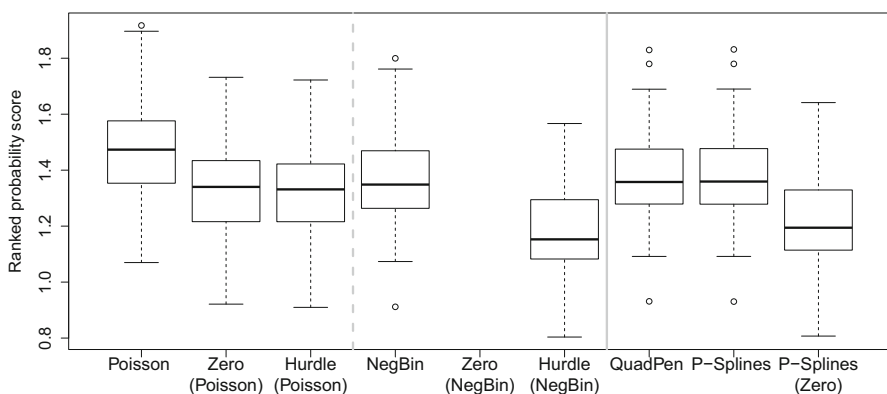

Fig. 11 Analysis of the boating trips data. The boxplots show the ranked probability score (for $r \in \{0, \ldots, 30\}$) of the six classical models (left and middle panel) and the three transition models (right panel). Due to computational problems, the zero-inflated Negative Binomial model was excluded from the analysis. All methods were fitted to 100 subsamples without replacement of size 438 and evaluated on the remaining 219 observations each

accounting for excess zeros strongly improves the predictive performance within all three model classes. The clearly best-performing models (on average), were the Hurdle Negative Binomial model (sixth boxplot) and the extended transition model (ninth boxplot). This again underlines that the proposed model flexibly adapts to the data and is not inferior to tailored classical models in terms of prediction.

The results from fitting the two superior models to the whole sample of $n = 657$ individuals are given in Table 4. Both models closely coincide. A high quality ranking increases the probability for outcome values greater than zero ($\hat{\beta}_0 = 1.428/1.495$). Within the second part of the model the expected number of boating trips was significantly higher for individuals that (i) did water-skiing at the lake, and (ii) payed a user fee. On the other hand, the expected number of boating trips decreased with the expenditure spent when visiting the lake. In terms of the transition model, the continuation ratio (7) is decreased by the factor $\exp(-0.9) = 0.407$ with an increased expenditure of 100 USD, since the corresponding parameter is -0.009.

7 Software

A crucial advantage of the proposed transition models is that they can be fitted using standard software for binary response models. Before fitting models one has to generate the binary data $(\tilde{Y}_{i0}, \tilde{Y}_{i1}, \ldots, \tilde{Y}_{iY_i})^T = (1, 1, \ldots, 1, 0)$ that encode the transitions up to the observed outcome. This is done by the generation of an

Table 4 Analysis of the boating trips data

	Zero			Non-Zero		
	Coef	SE	z-value	Coef	SE	z-value
Hurdle (NegBin)						
Quality	1.428	0.118	12.143	0.156	0.103	1.524
Ski	0.216	0.374	0.579	0.705	0.260	2.711
Income	0.012	0.095	0.124	− 0.087	0.088	− 0.987
Userfee	–	–	–	1.113	0.526	2.115
Cost	− 0.004	0.003	− 1.207	− 0.010	0.004	− 2.847
P-Splines (Zero)						
Quality	1.495	0.101	14.821	0.110	0.062	1.782
Ski	0.188	0.315	0.595	0.393	0.161	2.443
Income	− 0.033	0.083	− 0.398	− 0.068	0.053	− 1.299
Userfee	–	–	–	1.021	0.317	3.216
Cost	− 0.003	0.003	− 1.085	− 0.009	0.002	-3.963

Parameter estimates, standard errors and z-values obtained from fitting the Hurdle Negative Binomial model (upper panel) and the extended transition model (lower panel) with penalized B-splines accounting for excess zeros

augmented data matrix, which is composed of a set of smaller (augmented) data matrices for each individual. The resulting matrix has $\sum_{i=1}^{n} Y_i$ rows. In R the augmented data matrix can be generated using the function `dataLong()` in the R package **discSurv** (Welchowski and Schmid 2019). Estimates of the model with a quadratic penalty on the intercepts can be computed using the function `ordSmooth()` in the R package **ordPens** (Gertheiss 2015), estimates of the model with penalized B-splines can be obtained using the function `gam()` in the R package **mgcv** (Wood 2006).

8 Concluding remarks

A semiparametric alternative for the modeling of count data is proposed. The models are very flexible, as they do not assume a fixed distribution for the response variable, but adapt the distribution to the data by using smoothly varying coefficients. The extended form of the model further allows that also the regression coefficients vary smoothly over categories. Importantly, the models also directly account for the presence of excess zeros This has the key advantages that no parametric distribution has to be chosen and no specific two-component model needs to be built.

Our simulations and applications showed that in terms of prediction (measured by the ranked probability score) the transition model performs at least as good as the classical models in various settings. In more complex scenarios (as considered in simulation (e)) the transition model is even superior to its parametric alternatives. It was also illustrated that the extension to varying regression coefficients (in the third application) further enhances the flexibility and the predictive ability of the model class.

An important advantage of the transition models is that they can be embedded into the class of binary regression models. Therefore all inference techniques including asymptotic results to obtain confidence intervals that have been shown to hold for this class of models can be used. A further consequence is that selection of covariates can be done within that framework. Selection of covariates may be demanding even for a moderate number of covariates, because the number of regression coefficients highly depends on the number of response categories. For example, regularization methods as the lasso are applicable but beyond the scope of this article.

We restricted our consideration to the logistic link function. Although it is an attractive choice, because of the simple interpretation of effects, it should be noted that also alternative link function (e.g., the complementary log-log link function) can be used for fitting. Also the assumption of linear predictors can be relaxed by additive predictors, for example, by the use of spline functions for continuous covariates. Then one might investigate non-linear (smoothly varying) coefficients.

References

Agresti A (2002) Categorical data analysis. Wiley, New York

Aitkin M (1979) The analysis of unbalanced cross-classifications. J R Stat Soc Ser A (General) 142:404–404

Bocci C, Grassini L, Rocco E (2020) A multiple inflated negative binomial hurdle regression model: analysis of the Italians' tourism behaviour during the great recession. Stat Methods Appl. https://doi.org/10.1007/s10260-020-00542-6

Böhning D, van der Heijden PGM (2009) A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. Ann Appl Stat 3:595–610

Cameron AC, Trivedi PK (2005) Microeconometrics: methods and applications. Cambridge University Press, Cambridge

Cameron AC, Trivedi PK (2013) Regression analysis of count data, vol 53. Cambridge University Press, Cambridge

Consul PC (1998) Generalized Poisson distributions. Marcel Dekker, New York

Creel MD, Loomis JB (1990) Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California. J Agric Econ 72:434–441

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. Stat Modell 7(2):125–153

Czado C, Gneiting T, Held L (2009) Predictive model assessment for count data. Biometrics 65(4):1254–1261

Deb P, Trivedi PK (1997) Demand for medical care by the elderly: a finite mixture approach. J Appl Econ 12(3):313–336

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and Penalties. Stat Sci 11:89–121

Fahrmeir L, Kaufmann H (1987) Regression model for nonstationary categorical time series. J Time Ser Anal 8:147–160

Fahrmeir L, Tutz G (2001) Multivariate statistical modelling based on generalized linear models. Springer, New York

Famoye F, Singh KP (2003) On inflated generalized Poisson regression models. Adv Appl Stat 3(2):145–158

Famoye F, Singh KP (2006) Zero-inflated generalized Poisson model with an application to domestic violence data. J Data Sci 4(1):117–130

Gertheiss J (2015) ordPens: selection and/or smoothing of ordinal predictors. R package version 0.3-1. https://CRAN.R-project.org/package=ordPens

Giles D (2007) Modeling inflated count data. In: MODSIM 2007 international congress on modelling and simulation. Modelling and Simulation Society of Australia and New Zealand, Christchurch, NZ, pp 919–925

Gneiting T, Raftery A (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–376

Greene WH (1994) Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. NYU working paper No EC-94-10

Gschoessl S, Czado C (2006) Modelling count data with overdispersion and spatial effects. Stat Pap 49(3):531–552

Gupta PL, Gupta RC, Tripathi RC (2004) Score test for zero inflated generalized Poisson regression model. Commun Stat Theory Methods 33(1):47–64

Hayat M, Higgins M (2014) Understanding poisson regression. J Nurs Educ 53:207–215

Hilbe J (2011) Negative binomial regression. Cambridge University Press, Cambridge

Hilbe J (2014) Modeling count data. Cambridge University Press, Cambridge

Joe H, Zhu R (2005) Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. Biom J 47(2):219–229

Kaufmann H (1987) Regression models for nonstationary categorical time series: asymptotic estimation theory. Ann Stat 15:79–98

Kedem B, Fokianos K (2002) Regression models for time series analysis. Wiley, New York

Kleiber C, Zeileis A (2008) Applied econometrics with R. Springer, New York

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

Li CS et al (2012) Identifiability of zero-inflated Poisson models. Braz J Prob Stat 26(3):306–312

Loeys T, Moerkerke B, De Smet O, Buysse A (2012) The analysis of zero-inflated count data: beyond zero-inflated Poisson regression. Br J Math Stat Psychol 65:163–180

Maxwell O, Mayowa BA, Chinedu IU, Peace AE (2018) Modelling count data; a generalized linear model framework. Am J Math Stat 8:179–183

McCullagh P (1980) Regression model for ordinal data (with discussion). J R Stat Soc B 42(2):109–127

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, New York

Min A, Czado C (2010) Testing for zero-modification in count regression models. Stat Sin 20:323–341

Mullahy J (1986) Specification and testing of some modified count data models. J Econ 33(3):341–365

Mullahy J (1986b) Specification and testing of some modified count data models. J Econ 33:341–365

Nikoloulopoulos A, Karlis D (2008) On modeling count data: a comparison of some well-known discrete distributions. J Stat Comput Simul 78:437–457

Ozuna T, Gomez IA (1995) Specification and testing of count data recreation demand functions. Empir Econ 20(3):543–550

Payne EH, Hardin JW, Egede LE, Ramakrishnan V, Selassie A, Gebregziabher M (2017) Approaches for dealing with various sources of overdispersion in modeling count data: scale adjustment versus modeling. Stat Methods Med Res 26:1802–1823

Pohlmeier W, Ulrich V (1995) An econometric model of the two-part decisionmaking process in the demand for health care. J Hum Resour 30:339–361

Rigby R, Stasinopoulos D, Akantziliotou C (2008) A framework for modelling overdispersed count data, including the poisson-shifted generalized inverse gaussian distribution. Comput Stat Data Anal 53:381–393

Sellers K, Shmueli G (2010) A flexible regression model for count data. Ann Appl Stat 4:943–961

Tutz G (2012) Regression for categorical data. Cambridge University Press, Cambridge

Tutz G, Schmid M (2016) Modeling discrete time-to-event data. Springer, New York

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York

Welchowski T, Schmid M (2019) discSurv: discrete time survival analysis. R package version 1.4.1. http://CRAN.R-project.org/package=discSurv,

Wood SN (2006) Generalized additive models: an introduction with R. Chapman & Hall/CRC, London

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

Zou Y, Geedipally SR, Lord D (2013) Evaluating the double Poisson generalized linear model. Accid Anal Prev 59:497–505