Baetschmann, Gregori; Winkelmann, Rainer

Working Paper

A dynamic hurdle model for zero-inflated count data: With an application to health care utilization

SOEPpapers on Multidisciplinary Panel Data Research, No. 648

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Baetschmann, Gregori; Winkelmann, Rainer (2014) : A dynamic hurdle model for zero-inflated count data: With an application to health care utilization, SOEPpapers on Multidisciplinary Panel Data Research, No. 648, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/96500

SOEP Das Sozio-oekonomische Panel

648 2014

SOEPpapers
on Multidisciplinary Panel Data Research

SOEP – The German Socio-Economic Panel Study at DIW Berlin 648-2014

A Dynamic Hurdle Model for Zero-Inflated Count Data: With an Application to Health Care Utilization

Gregori Baetschmann and Rainer Winkelmann

DIW SOEP

SOEPpapers on Multidisciplinary Panel Data Research
at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | soeppapers@diw.de

A Dynamic Hurdle Model for Zero-Inflated Count Data: With an Application to Health Care Utilization*

Gregori Baetschmann
University of Zurich

Rainer Winkelmann
University of Zurich and IZA

March 2014

Abstract

Excess zeros are encountered in many empirical count data applications. We provide a new explanation of extra zeros, related to the underlying stochastic process that generates events. The process has two rates, a lower rate until the first event, and a higher one thereafter. We derive the corresponding distribution of the number of events during a fixed period and extend it to account for observed and unobserved heterogeneity. An application to the socio-economic determinants of the individual number of doctor visits in Germany illustrates the usefulness of the new approach.

JEL Classification: C25, I10
Keywords: excess zeros, Poisson process, exposure, hurdle model

*Corresponding author: Rainer Winkelmann, University of Zurich, Department of Economics, Zurichbergstrasse 14, CH-8032 Zurich; email: rainer.winkelmann@econ.uzh.ch. We are grateful to Robert Jung, Johannes Kunz, Joao Santos Silva and Kevin Staub for valuable comments, and to Daniel Auer for very able research assistance.

1 Introduction

A large literature on the analysis of count data is now available (Winkelmann 2008, Cameron and Trivedi 2013), but only a small portion of it deals with the specialness of zeros. If counts are the outcome of individual choice, zeros can result from a corner solution in the optimization, and this aspect should be addressed by the econometric model. In practice, zeros often deserve special attention because of their ubiquity: when analyzing the number of job changes by a worker during a decade, the annual number of doctor visits or the weekly number of purchases of a specific good at a given supermarket, there are often many more zeros in the data than predicted by a standard count data model, a situation denoted as zero-inflation, or "excess zeros".

The existing zero-inflation approaches in the literature have focused on unobserved heterogeneity and two-part models. An extreme form of unobserved heterogeneity is obtained in a finite mixture model with two types of observation units where one type never experiences the event (leading to a count of zero) and the other type has a standard count distribution (Mullahy 1986, Lambert 1992). In a hurdle model, like in a two-part model for continuous responses used in health economics, a binary model for the 0/1+ decision is combined with a truncated-at-zero count data model for positive outcomes (Mullahy 1986). Both approaches have been regularily used in applied work, examples including Pohlmeier and Ulrich (1995), Street, Jones and Furuta (1999), Campolieti (2002), Winkelmann (2008) and Sari (2009).

In this paper, we propose a new, alternative model for zero-inflation, based on a dynamic hurdle specification. Starting point is a stochastic process for the timing of events. The Poisson process for the number of events prior to time T requires that times between events are i.i.d. exponential with constant hazard rate λ. We define a generalization, where the distribution of the time to first event can have a different rate than the distribution of times between subsequent events. In the taxonomy of Heckman and Borjas (1980), our model allows for occurrence dependence but rules out duration dependence. An extended version of

the model also accounts for unobserved heterogeneity. In statistics and biometrics, generalizations of the Poisson process such as the one explored here fall into the class of birth process models (see e.g. Janardan, 1980, Faddy, 1997).

Occurrence dependence leads to zero-inflation if the rate is low until the first event and higher thereafter. In contrast to the static hurdle model, the higher rate only applies to the time left between the first event and T. The model thus adds a dynamic selection effect: variation in the first rate systematically affects the expected arrival time of the first event, and hence the duration for which the process is in the second state. As a consequence, the probability of a zero and the distribution of positive outcomes are not independent.

Our dynamic hurdle model has a number of useful properties. It nests the standard Poisson (or negative binomial) model so that testing for the absence of occurrence dependence is straightforward. The parameters are easy to interpret, as they are hazard rates of the time to first event and time between further events, respectively. The mean is available in closed form, so that simple analytical formulas for marginal effects and average treatment effects exist. Because the model is based on a fully specified structural stochastic process, it is simple to incorporate varying time of exposure (Baetschmann and Winkelmann, 2013). The methods that we develop in this article are designed for cross-sectional count data. With panel data or multiple-spell duration data, other approaches would be feasible and occurrence dependence could be tested more directly.

The paper proceeds as follows. In the next section, we present the standard models for count dependent variables as well as the static hurdle model. In section 3, we derive the dynamic hurdle model and discuss its properties, including some possible specification tests. The new model is used, in section 4, to estimate the socio-economic determinants of the number of quarterly visits to a physician, based on survey data from the German Socio-Economic Panel for the year 2006. Section 5 concludes.

2 Modeling count data

It is well known that if events occur randomly over time, without occurrence dependence, duration dependence, or unobserved heterogeneity, the number of events during a unit time interval is Poisson distributed with probability function

$$\Pr(Y = k) = \frac{\exp(-\lambda)\lambda^k}{k!}, \;\; k = 0, 1, 2, \ldots \tag{1}$$

where λ is the constant rate, or intensity, of the process and also the mean of the Poisson distribution. Violation of randomness or homogeneity lead to different count data models ("non-Poissonness"). In the past, a considerable amount of research has been devoted to the consequences of unobserved heterogeneity (e.g., Hausman, Hall and Griliches, 1984, Cameron and Trivedi, 1986) and duration dependence (e.g., Winkelmann, 1995, McShane et al., 2008). If λ follows a gamma distribution, the resulting marginal probability function for Y is negative binomial. The negative binomial distribution has, for a given mean, a larger variance than the Poisson distribution (overdispersion). It also has a higher probability of a zero. Similarly, extra zeros can be generated from a model where the times between events have a distribution with negative duration dependence (Winkelmann, 1995).

In many applications, extra zeros (relative to the Poisson model) generated by the above models are insufficient to account for the full amount of zeros in the data. All single index models have to compromise between the large proportion of zeros, which tends to lower the mean, and a right-skewed distribution of counts with large non-zero values, which tends to increase it. Moreover, one often has a substantive interest in treating the zero-generating process separately from the process for strictly positive outcomes, which requires different sets of parameters. In the fixed-hurdle (FH) count data model (Mullahy, 1986)

$$\Pr(Y = k) = \begin{cases} \phi & \text{for } k = 0 \\ (1 - \phi)\dfrac{f(k)}{1 - f(0)} & \text{for } k \geq 1 \end{cases} \tag{2}$$

where $f(k)$ denotes the probability function of a standard count data model, usually either the Poisson or the negative binomial distribution. Thus, the distribution of positive counts depends on two factors, the probability of crossing the hurdle and the conditional-on-positives distribution. In health services research, $1-\phi$ is known as the utilization probability, i.e. the probability of using services at least once. Often, ϕ is specified as the survivor function of the exponential distribution. When $1-\phi > f(0)$, the data are zero-inflated relative to the base distribution. The distribution $f(k)$ models the frequency of repeat visits for $k \geq 1$. Pohlmeier and Ulrich (1995) argue that such a hurdle model can well represent actual decision processes, for example in the context of the demand for doctor visits, where a first decision to contact a general practitioner might be followed by a number of re-appointments or referrals to specialists that are subject to a different mechanism.

While this approach generates a kind of "occurrence dependence", it is not derived from an underlying stochastic process. It therefore ignores the timing dimension, i.e., the difference it makes whether the first contact was made earlier or later during the observation period. Our new model, by contrast, directly addresses the dynamic hurdle selection. The process switches from a low rate for the first occurrence (state 1) to a higher rate for subsequent occurrences (state 2) or vice versa. The hurdle is dynamic, since the timing of the hurdle-crossing from state 1 to state 2 is endogenously determined. The key difference to the fixed hurdle model is the time effect: the lower the stage 1 rate, the later the expected time of crossing and the less time is spent in the state 2 process. This time effect is ignored by the fixed hurdle model.

3 Dynamic hurdle count models

3.1 Timing of the first event

There is a fundamental relationship between the time of the first event, ϑ_1, and the total number of events between 0 and T, $Y(0,T)$. Suppose that a first event occurs at $\vartheta_1 = t$ and $k-1$ events occur between t and T, then, for $k \geq 1$, the total number of events between 0 and T is equal to k. Therefore

$$\begin{aligned} \Pr[Y(0,T) = k, \vartheta_1 = t] &= \Pr[Y(t,T) = k-1, \vartheta_1 = t] \\ &= \Pr[Y(t,T) = k-1] f_1(t) \end{aligned} \tag{3}$$

where $f_1(t)$ is the density function of the time of the first event. Moreover, the marginal probability function $\Pr[Y(0,T) = k]$ can be obtained as

$$\Pr[Y(0,T) = k] = \begin{cases} \displaystyle\int_0^T \Pr[Y(t,T) = k-1] f_1(t) dt & \text{if } k \geq 1 \\ \Pr(\vartheta_1 > T) & \text{if } k = 0 \end{cases}$$

with mean

$$\begin{aligned} \mathrm{E}(Y(0,T)) &= \sum_{k=1}^{\infty} k \int_0^T \Pr[Y(t,T) = k-1] f_1(t) dt \\ &= \int_0^T \sum_{k=0}^{\infty} (k+1) \Pr[Y(t,T) = k] f_1(t) dt \\ &= \Pr(y > 0) + \mathrm{E}_t[\mathrm{E}Y(t,T)] \end{aligned} \tag{4}$$

These are completely general results obtained without imposing any particular structure on the stochastic process. Even so, they tell us that the mean of a count variable is equal to the probability of a positive count plus the expected number of events occuring after the time of the first event, i.e., after crossing the dynamic hurdle.

3.2 A dynamic hurdle Poisson model

To obtain the dynamic hurdle Poisson model, we make the following assumptions. First, $f_1(t)$ is the density of the exponential distribution with rate λ_1. Second, events in the second state are generated from a Poisson process with rate λ_2, such that $Y(t,T) \sim Poisson(\lambda_2(T-t))$. With these assumptions, for $k \geq 1$,

$$\Pr[Y(0,T)=k] = \int_0^T \frac{\exp(-\lambda_2(T-t))[\lambda_2(T-t)]^{k-1}}{(k-1)!}\lambda_1 \exp(-\lambda_1 t)dt \tag{5}$$

From now on, we use the normalization $T=1$. One can show (see Appendix A) that the integral has closed form solution, and for $y \geq 1$, the probability function of the dynamic hurdle Poisson model is given by

$$f_{DHP}(y;\lambda_1,\lambda_2) = \frac{\lambda_1 \lambda_2^{y-1}\exp(-\lambda_1)}{(\lambda_2-\lambda_1)^y}\left[1 - \sum_{j=0}^{y-1}\frac{\exp(-(\lambda_2-\lambda_1))(\lambda_2-\lambda_1)^j}{j!}\right]$$

It simplifies to that of the Poisson distribution if $\lambda_1 = \lambda_2$.

3.3 Properties

For the dynamic hurdle Poisson model, the expected time spent in the first state is

$$\begin{aligned}
\mathrm{E}(t;\lambda_1) &= \Pr(t \geq 1;\lambda_1) + \Pr(t<1;\lambda_1)\mathrm{E}(t|t<1;\lambda_1) \\
&= \Pr(y=0) + \int_0^1 \exp(-\lambda_1 t)\lambda_1 t\, dt \\
&= \exp(-\lambda_1) + \lambda_1^{-1} - (1+\lambda_1^{-1})\exp(-\lambda_1) \\
&= \lambda_1^{-1}(1-\exp(-\lambda_1))
\end{aligned}$$

and the expected time spent in the second state is therefore

$$\mathrm{E}(1-t;\lambda_1) = 1 - \lambda_1^{-1}(1-\exp(-\lambda_1)) \tag{6}$$

We can use (6) to rewrite (4) as

$$\begin{aligned} \mathrm{E}_{DHP}(y;\lambda_1,\lambda_2) &= \Pr(y>0;\lambda_1) + \lambda_2 \mathrm{E}(1-t;\lambda_1) \\ &= [1-\exp(-\lambda_1)] + \lambda_2[1-\lambda_1^{-1}(1-\exp(-\lambda_1))] \\ &= \lambda_2 + (1-\lambda_2/\lambda_1)[1-\exp(-\lambda_1)] \end{aligned} \tag{7}$$

As required, the expected value reduces to the Poisson mean when $\lambda_1 = \lambda_2$. The expected value is greater than λ_2 when $\lambda_1 > \lambda_2$, and smaller otherwise.

3.4 Comparison to the fixed hurdle Poisson model

Equation (7) illustrates an important property of the dynamic hurdle Poisson model. The expectation is the sum of the probability of passing the dynamic hurdle, plus the state 2 rate times the expected duration in state 2. Thus λ_1 affects the overall mean through two separate channels. First, it affects the probability of crossing the hurdle, and second, it affects the expected duration spent in the second state. This distinction is absent in the fixed hurdle Poisson model, where the expectation is given by

$$\begin{aligned} \mathrm{E}_{FHP}(y;\lambda_1,\lambda_2) &= \Pr(y>0;\lambda_1)\mathrm{E}(y|y>0;\lambda_2) \\ &= [1-\exp(-\lambda_1)]\,\frac{\lambda_2}{1-\exp(-\lambda_2)} \end{aligned} \tag{8}$$

One can show that $\partial \mathrm{E}_{DHP}(y)/\partial\lambda_1 > \partial \mathrm{E}_{FHP}(y)/\partial\lambda_1$ if the two models have the same expected value and the same fraction of zeros. On the other hand, the probability of a zero is identical in the two models,

$$\Pr_{DHP}(0|\lambda_1,\lambda_2) = \Pr_{FHP}(0|\lambda_1,\lambda_2) = \exp(-\lambda_1)$$

and zero-inflation therefore arises in either case whenever $\lambda_1 < \lambda_2$.

3.5 Observed heterogeneity

In cross-sectional count data applications, we observe independent pairs of observations (y_i, x_i), $i = 1, \ldots, n$, and the interest usually centers on the effect of covariates on the conditional mean $\mathrm{E}(y_i|x_i)$, or some other feature of the conditional distribution of $f(y_i|x_i)$. The standard way of introducing covariates is to let $\lambda_{ij} = \exp(x_i'\beta_j)$, $j = 1, 2$, where x_i denotes the $(k \times 1)$-vector of covariates, including a constant, and β the conformable parameter vector. This parameterization ensures positive rates and implies a semi-elasticity interpretation for β. Further it allows to treat exposure T_i, the length of the observation period, as a standard covariate. Incorporating exposure explicitly in the model is necessary if T_i varies between individuals and cannot be normalized to one therefore.

3.6 Decomposing the mean effect

The FH model (see 8) has a standard two-part structure, where the two parts are independent. This allows for a straightforward decomposition of the overall effect into an effect at the extensive margin and an effect at the intensive margin:

$$\frac{\partial \mathrm{E}_{FH}(y; \lambda_1(x), \lambda_2(x))}{\partial x} = \frac{\partial \Pr(y > 0; \lambda_1(x))}{\partial x} \mathrm{E}(y|y > 0; \lambda_2(x)) + \frac{\partial \mathrm{E}(y|y > 0; \lambda_2(x))}{\partial x} \Pr(y > 0; \lambda_1(x)) \qquad (9)$$

It is useful to think of the extensive margin effect as a participation effect (i.e., whether or not one has seen a doctor, or changed a job at all), whereas the intensive margin effect is the effect for participants, also called the conditional-on-positives effect. Note that the extensive margin effect is the change in the probability of participation *times the average outcome of participants.* This may overstate the true causal effect if those at the margin of participation have below average outcomes once they participate (Staub,

2013).

The dynamic hurdle model lends itself to a more detailed decomposition of marginal mean effects. Differentiating (7) with respect to x, the DH model implies the following decomposition of the partial derivative of the overall mean:

$$\frac{\partial \mathrm{E}_{DH}(y;\lambda_1(x),\lambda_2(x))}{\partial x} = \frac{\partial \Pr(y>0;\lambda_1(x))}{\partial x} + \lambda_2 \frac{\partial \mathrm{E}(T-t;\lambda_1(x))}{\partial x} + \mathrm{E}(T-t;\lambda_1(x))\frac{\partial \lambda_2(x)}{\partial x} \tag{10}$$

Here, the extensive margin effect is the change in the participation probability, multiplied by one, and hence always smaller than the effect under the standard two-part decomposition. The reason is that the marginal observation does not spend any time in the state 2 process, and hence at the margin gets a weight of $\mathrm{E}(y|y>0, 1-t=0)=1$. Also note, that the intensive margin effect can now be further decomposed into a time effect and a productivity effect.

3.7 Unobserved heterogeneity

Suppose that $f(y_i|x_i,\epsilon_i)$ is Poisson distributed with rate $\exp(x_i'\beta+\epsilon_i)$ where $\exp(\epsilon_i)\equiv u_i>0$ captures the effect of unobserved variables z_i. It is well known that if u_i is i.i.d. gamma distributed with mean 1 and variance α, the distribution of y_i conditional on x_i but unconditional on u_i is negative binomial with mean λ and variance $\lambda(1+\lambda\alpha)$. The DHP model can be extended along the same lines. Let the two rates be given by $\lambda_j(x_i,u_i)=\exp(x_i'\beta_j)u_i$, $j=1,2$, where u_i is a gamma distributed individual heterogeneity term that equally affects both rates of the DHP model. The conditional probability of observing a count y is then $f_{DHP}(y;\lambda_1 u,\lambda_2 u)$, and integrating over the unobserved u leads to the unconditional probability function

$$\begin{aligned} & f_{DHNB}(y;\lambda_1,\lambda_2,\alpha) \\ & \quad = \int_0^\infty f_{DHP}(y;\lambda_1 u,\lambda_2 u)g(u;\alpha)du \end{aligned} \tag{11}$$

$$= \begin{cases} (\lambda_1/\alpha+1)^{-\alpha} & \text{for y=0} \\ \dfrac{\lambda_1\lambda_2^{y-1}}{(\lambda_2-\lambda_1)^y}\left(\dfrac{\alpha}{\alpha+\lambda_1}\right)^{\alpha}\left[1-\sum_{j=0}^{y-1}(1-\theta)^j\,\theta^{\alpha}\dfrac{\Gamma(\alpha+j)}{\Gamma(\alpha)\Gamma(j+1)}\right] & \text{for } y=1,\,2,\,3,\ldots, \end{cases}$$

where $\theta = (\alpha+\lambda_1)/(\alpha+\lambda_2)$ and $g(u;\alpha)$ is the gamma density function. For $\lambda_2 > \lambda_1$, the term in squared brackets is equal to the complementary cumulative distribution function of the negative binomial distribution. The mean of the dynamic hurdle negative binomial model is given by

$$\begin{aligned} \mathrm{E}_{DHNB}(y|\lambda_1,\lambda_2,\alpha) &= \int_0^{\infty} \lambda_2 u + (1-\lambda_2 u/\lambda_1 u)\,(1-\exp(-\lambda_1 u))g(u;\alpha)du \\ &= \lambda_2 + (1-\lambda_2/\lambda_1)(1-f_{NB}(0;\lambda_1,\alpha)) \\ &= \lambda_2 \mathrm{E}(1-t,\lambda_1,\alpha) + \Pr(y>0;\lambda_1,\alpha) \end{aligned} \tag{12}$$

It preserves the essential structure of the mean of the dynamic hurdle Poisson model, and simplifies to it for $\alpha = 0$.

3.8 Estimation and testing

One can estimate β_1, β_2 and α by maximum likelihood (Stata code is available from the authors upon request). In empirical applications, the interest is often in testing for the presence of excess zeros. Under the null of no additional zeros, $\lambda_1 = \lambda_2$ which requires that $\beta_1 = \beta_2$. The DH Poisson model simplifies to a Poisson model, and the DH negative binomial model to a negative binomial model, respectively. The likelihood ratio test statistic is chi-squared distributed with k degrees of freedon. A similar test is possible for the FH model.

Since the DH, FH, and ZIP models are not pairwise nested, one can use the Vuong-Test (Vuong, 1989) for overlapping models to discriminate between them. The two-step procedure requires first testing for equivalence. For example, in the case of the DH and FH models, both nest the Poisson model and thus

are equal in that case. The ZIP model can be rewritten as a hurdle model with utilization probability $\phi = p + (1-p)f(0)$ where p is the probability of an extra zero. The two are thus equivalent in the constant-only case. Once these conditions for equivalence are rejected, the second stage of the test determines whether one of the two models significantly outperforms the other in terms of minimizing the Kullback-Leibler distance (see Vuong, 1989, for additional detail). Alternatively, one can select the best model using an information criterion. However, since the models have an identical number of parameters, an adjustment for degree of freedom is not required, and the model with the highest log likelihood value will be selected.

Finally, there is a possibility of an informal specification test of the DH model. Define the binary event "positive count yes/no". Under the assumption of the DH model, this event has Bernoulli distribution with complementary log-log link and parameter λ_1. Thus, λ_1 is identified from a separate binary model and does not require estimation of the full DH model. An informal specification test can be based on a comparison of $\hat{\beta}_1$ in the full DH model with that of a simple binary model (i.e., the first stage of the FH model). If the two differ a lot, the DH specification is likely wrong.

4 Application: the socio-economic determinants of the number of doctor visits in Germany

We apply the new dynamic hurdle models to estimate the socio-economic determinants of the quarterly number of doctor visits in Germany. Most people living in Germany are covered by statutory health insurance which is financed through payroll deductions. Services are offered at no cost but there is a co-payment for prescription drugs. In addition, between 2004 and 2012, a one-time quarterly fee of 10 Euros had to be paid for the first visit to the family doctor, or general practitioner ("Praxisgebühr"). Further visits, or subsequent referrals to a specialist, were then free of charge. Effectively, this non-linear pricing

scheme mirrors to the structure of the dynamic hurdle process developed in this paper: in state 1, without a previous visit during the quarter, the price of a visit is 10 Euro. Thereafter, in state 2, it drops to 0. Provided people respond to price incentives, the state 1 rate should be lower than the state 2 rate, leading to zero inflation in the distribution of the quarterly number of visits. Although we do not observe the timing of the first visit, if any, the dynamic hurdle model can account for this latent selection process.

The data for the analysis have been extracted from the Socio-Economic Panel (SOEP, see Wagner et. al, 2007). As our models are designed for cross-section data, we use data from a single wave, the year 2006. There are two reasons for choosing this particular year. First, by that time, people should have adapted their demand behavior to the new pricing system introduced two years earlier. And second, there are certain questions on interesting health behaviors that were included in the 2006 wave but are missing in other years.

The dependent variable is the self-reported number of doctor visits during the previous three months. While most interviews are conducted before the summer holidays, there is substantial variation by month and day of the month. Thus, the three months reporting period does not necessarily coincide with the calendar quarter that is relevant for the dynamic pricing of doctor visits. There are two possible responses to this reporting mismatch. First, following Farbmacher and Winter (2013), one can select a subset of persons who were interviewed within plus or minus 10 days of the end of the quarter (March 31, June 30, or September 30, respectively; there are no interviews held in the second half of December or at the beginning of January). This is what we do in our analysis. Second, it is easy to show that the expected duration spent in the low hazard state 1 (where the price of a visit is 10 Euros) unaffected by the reporting period, as is the expected duration spent in state 2 (where visits are free). One could thus estimate the DH model with the unrestricted sample as well.

— — — Table 1 about here — — —

After further limiting our sample to those covered by statutory health insurance, being older than 18 (those aged 18 or younger were exempt from the quarterly fee) and younger than 70, we are left with 2966 observations. Table 1 shows selected summary statistics of the variables employed in the estimation. The number of doctor visits has a mean of 2.1, and 34 percent of all persons in the sample did not visit a doctor during the previous quarter. The remaining variables are used as determinants of health care utilization: two demographic variables (age and its square as well as gender), two socio-economic variables (years of schooling and log of disposable household income), two indicators of health behaviors (current smoker and the body mass index), and a disability indicator. Thus, the hazard rate in each part of the model depends on a total of 9 parameters each.

We intended to estimate the fixed and dynamic hurdle models without and with unobserved heterogeneity. However, the DHNB model did not converge for our particular dataset. Convergence problems with modified negative binomial models are not unusual in our experience, including problems for the zero-truncated negative binomial routine available in STATA, that is a component of the fixed hurdle model with unobserved heterogeneity. For example, it may be the case that the true process displays underdispersion in the positive part of the model, but overdispersion overall, so that the mixture hurdle model will never converge.

As a consequence, we report in Table 2 estimates from three models, the standard Poisson model, the fixed hurdle Poisson model, and the dynamic hurdle Poisson model. Columns (2) and (3) show the results for the hazard rates λ_1 and λ_2 for the two states of the fixed hurdle Poisson model and columns (4) and (5) the corresponding results for the dynamic hurdle Poisson model, respectively. A positive β means that a unit increase in the associated regressor increases the rate of the state specific process by $[\exp(\beta) - 1] \times 100$ percent. In the case of the first state, an increase in the hazard rate implies a reduction in the probability of a zero count.

In terms of overall fit, the simple Poisson model is clearly rejected against either of the two generalizations, based on a likelihood ratio test. The DHP model has a substantially higher loglikelihood value than the FHP model (-5961.5 as compared to -6153.1). We also estimated the zero-inflated Poisson model (see Winkelmann, 2008, for specification details). While it offers an improvement over the simple Poisson and the FHP models, its loglikelihood of -6150.9 falls short of that of the DHP model. Thus, any of the standard model selection criteria would pick the DHP model, and the same is true when applying the Vuong test for overlapping models.

— — — Table 2 about here — — —

The estimated partial effects of the FHP and the DHP models are qualitatively similar. As mentioned above, the probability of a zero visit equals $\exp(-\lambda_1)$ in both models, so it is reassuring that the estimated coefficients are of similar magnitude. The λ_2 parameters tend to be larger in the DHP model. The DHP estimates thereby compensate for the fact that the second, or state 2, rate applies only for a fraction of the entire period, in contrast to the FHP model, where such a time effect is ruled out. It should also be noted that the standard errors of $\hat{\beta}_2$ are substantially higher in the DHP than in the FHP model. The dynamic interdependence between the two rates in the DHP model means that the second rate cannot be estimated independently of the first rate, reducing the precision of the estimator. On the other hand, $\hat{\beta}_1$ is estimated somewhat more precisely in the DH model than in the FH model, as it uses the distribution of the positives as an additional source of identification.

There are some interesting asymmetries between extensive margin (λ_1) and intensive margin (λ_2) effects in the DHP model. For instance, income has no effect at the extensive margin, but a statistically significant negative effect on the number of doctor visits at the intensive margin, where a 10 percent increase in income is predicted to reduce the hazard rate for each further doctor visit by 0.7 percent. The opposite pattern is observed for current smokers: perhaps surprisingly, the extensive margin effect is negative, while there

is no effect at the intensive margin. Individuals with disabilities have higher hazard rates in both states, and thus a higher predicted number of doctor visits, than others. Women have more visits than otherwise similar men.

For the DHP model, the average predicted state 1 hazard rate in the sample is 1.1, the average predicted state 2 hazard rate is 3.8; the predicted average duration spent in state 2 is 0.38, i.e., a bit more than one month. Applying (4), the predicted average number of doctor visits is 2.13, close to the sample average of 2.05. One should be careful, though, in interpreting the predicted increase in the hazard rate following the first visit as resulting from the non-linear pricing introduced by the first-visit fee. While the rate increase implied by the DHP model is compatible with such an explanation, it seems much too large to be resulting from a 10 Euro change in price, and more importantly, there are other potential explanations for the same effect.

First and foremost, the model does not allow for unobserved heterogeneity. It is easy to show that unobserved heterogeneity leads to excess zeros relative to the Poisson model, so that ignoring unobserved heterogeneity will reappear as state dependence in the DHP, with a lower initial rate. Second, the non-linear pricing argument focusses on the demand side, while there can be supply side effects as well. For instance, doctors have an incentive to reduce appointments towards the end of a quarter and rather have patients come back at the beginning of the next quarter (Schmitz, 2013). This is unrelated to the one-time fee, which was collected by physicians but entirely passed on to the insurance companies, but rather due to a fixed budget allocation, or remuneration cap, applied to a physician on a per-patient-and-quarter basis. However, note that this effect would work in the opposite direction and tend to lower the state 2 hazard.

Finally, this analysis also entirely abstracted from the dynamics of sickness spells, that obviously also can lead to clustered visits and a non-stationary stochastic process. With the limited amount of information provided by a simple count of visits, is is perhaps not too surprising that there are limits to what we can

learn from the data, and without further assumptions, one should interpret the changes between λ_1 and λ_2, i.e. the dynamic hurdle, as the combined, reduced form effect of a number of different underlying factors.

5 Concluding remarks

This article develops a new approach for the regression analysis of zero inflated and overdispersed count data. In our model, zero inflation is modeled by assuming that the counts are generated from a non-stationary stochastic process, where there is a one-time increase in the otherwise constant underlying hazard rate at the time of the first event. The timing of the first event is thus endogenously determined in such a dynamic hurdle count model. Regressors are allowed to differentially affect the two hazard rates before and after the first event, and the effect on the overall count has three distinct channels: the probability of zero, the expected duration of the second state, and the hazard rate in that second state. We derive a Poisson and a negative binomial version of the model. In principle, the model can account for zero-deflation and underdispersion as well, if the rate to the first event is higher than the subsequent rate.

We apply the new model to an analysis of individual level health-care usage in Germany. The dependent variable is the self-reported number of visits to a doctor during the previous calendar quarter. Our model implicitly accounts for an institutional feature of the German health care system, namely that users have to pay a fee for the first visit per quarter, but not for subsequent ones. If users are responsive to prices, such a system can contribute to a one-time change in the underlying rate. Our results show that the dynamic hurdle model fits the data substantially better than two existing approaches for zero-inflated count data. The results are easy to interpret, since regression coefficients are semi-elasticities of the two hazard rates, and they allow for a useful distinction between intensive and extensive margin effects. On the substantive side, we find no evidence for the hypothesis that unhealthy behaviors (smoking, body mass index) increase the number of doctor visits.

References

Baetschmann, G. and R. Winkelmann (2013), Modelling zero-inflated count data when exposure varies: with an application to tumor counts, *Biometrical Journal*, 55, 679-686.

Cameron, A.C. and P.K. Trivedi (1986), Economic models based on count data: Comparisons and applications of some estimators and tests, *Jounral of Applied Econometrics*, 1, 29-53.

Cameron, A.C. and P.K. Trivedi (2013), *Regression analysis of count data* (Vol. 53), Cambridge University Press.

Campolieti, M. (2002), The recurrence of occupational injuries: estimates from a zero inflated count model, *Applied Economics Letters*, 9, 595-600.

Hausman, J., B.H. Hall and Z. Griliches (1984), Econometric models for count data with an application to the patents-R&D relationship, *Econometrica*, 52, 909-938.

Heckman, J.J. and G.J. Borjas (1980), Does unemployment cause future unemployment? definitions, questions and answers from a continuous time model of heterogeneity and state dependence, *Economica*, 47, 247-283.

Farbmacher, H. and J. Winter (2013), Per-period co-payments and the demand for health care: evidence from survey and claims data, *Health Economics*, 22, 11111123.

Feller, W. (1977), *An Introduction to Probability Theory and Its Applications*, second edition, New York: John Wiley & Sons.

Faddy, M. (1997), Extended Poisson process modelling and analysis of count data, *Biometrical Journal*, 39, 431-440.

Janardan, K.G. (1980), A stochastic model for the study of oviposition evolution of the pest callosobruchus maculatus on mung beans, phaseolus aureus, *Math. Biosciences*, 50, 231-238.

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

Mullahy, J. (1986), Specification and testing in some modified count data models, *Journal of Econometrics*, 33, 341-365.

McShane, B., M. Adrian, E.T. Bradlow and P.S. Fader (2008), Count models based on Weibull interarrival times, *Journal of Business & Economic Statistics*, 26, 369-378.

Pizer, S.D., and J.C. Prentice (2011), Time is money: outpatient waiting times and health insurance choices of elderly veterans in the United States, *Journal of Health Economics*, 30, 626-636.

Pohlmeier, W. and V. Ulrich (1995), An econometric model of the two-part decisionmaking process in the demand for health care, *Journal of Human Resources*, 30, 339-361.

Sari, N. (2009), Physical inactivity and its impact on healthcare utilization, *Health Economics*, 18, 885-901.

Schmitz, H. (2013), Practice budgets and the patient mix of physicians - The effect of a remuneration system reform on health care utilization, *Journal of Health Economics*, 32, 1240-1249.

Staub, K. (2013), A causal interpretation of extensive and intensive margin effects in generalized Tobit models, *Review of Economics and Statistics*, forthcoming.

Street, A., A. Jones and A. Furuta (1999), Cost sharing and pharmaceutical utilisation and expenditure in Russia, *Journal of Health Economics*, 18, 459-472.

Vuong, Q.H. (1989), Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, 57, 307-333.

Wagner, G.G., J.R. Frick and J. Schupp (2007), The German Socio-Economic Panel Study (SOEP) scope, evolution and enhancements, *Schmollers Jahrbuch*, 127 , 139-169.

Winkelmann, R. (1995), Duration dependence and dispersion in count-data models, *Journal of Business & Economic Statistics*, 13, 467-74.

Winkelmann, R. (2008), *Econometric Analysis of Count Data*, fifth edition, Berlin: Springer.

Appendix A. Derivation of the probability function of the stochastic hurdle model

The probability of a zero in the DH model equals the probability of a zero in a Poisson model with rate λ_1. If, $\lambda_1 = \lambda_2$ the DH model degenerates to a Poisson model. For $k = 1, 2, 3, \ldots$ and $\lambda_1 \neq \lambda_2$:

$$
\begin{aligned}
&\Pr(\quad Y = k|\lambda_1, \lambda_2) \\
&\qquad = \qquad \int_0^T \exp(-\lambda_1 t)\lambda_1 \exp(-\lambda_2(T-t))(\lambda_2(T-t))^{k-1}/(k-1)!dt \\
&\qquad = \qquad \lambda_1 \lambda_2^{k-1} \exp(-\lambda_2) \int_0^T \frac{\exp(\lambda_2 - \lambda_1)^t (T-t)^{k-1}}{(\lambda_2 - \lambda_1)(k-1)!} dt \\
&\qquad = \qquad \lambda_1 \lambda_2^{k-1} \exp(-\lambda_2) \left(\frac{\exp(\lambda_2 - \lambda_1)^t (T-t)^{k-1}}{(\lambda_2 - \lambda_1)(k-1)!} + \int_0^T \frac{\exp(\lambda_2 - \lambda_1)^t (T-t)^{k-2}}{(\lambda_2 - \lambda_1)(k-2)!} dt \right) \\
&\qquad = \qquad \frac{\lambda_2}{\lambda_2 - \lambda_1} \Pr(Y = k-1|\lambda_1, \lambda_2) - \frac{\lambda_1 \lambda_2^{k-1}}{\lambda_2 - \lambda_1} \exp(-\lambda_2)/(k-1)!
\end{aligned}
$$

Setting $T = 1$, and solving the recursive equation $p_k = \alpha p_{k-1} + c_k$ leads to:

$$
\begin{aligned}
&\Pr(\quad Y = k|\lambda_1, \lambda_2) \\
&\qquad = \qquad \alpha^{k-1} \Pr(k = 1|\lambda_1, \lambda_2) + \sum_{j=0}^{k-2} \alpha^j c_{k-j} \\
&\qquad = \qquad \frac{\lambda_1 \lambda_2^{k-1}}{(\lambda_2 - \lambda_1)^k} (\exp(-\lambda_1) - \exp(-\lambda_2)) - \sum_{j=0}^{k-2} \left(\frac{\lambda_2}{\lambda_2 - \lambda_1} \right)^j \frac{\lambda_1 \lambda_2^{k-j-1}}{\lambda_2 - \lambda_1} \frac{\exp(\lambda_2)}{(k-j-1)!} \\
&\qquad = \qquad \frac{\lambda_1 \lambda_2^{k-1}}{(\lambda_2 - \lambda_1)^k} \left(\exp(-\lambda_1) - \exp(-\lambda_2) \sum_{j=0}^{k-1} \frac{(\lambda_2 - \lambda_1)^j}{j!} \right) \\
&\qquad = \qquad \frac{\lambda_1 \lambda_2^{k-1} \exp(-\lambda_1)}{(\lambda_2 - \lambda_1)^k} \left(1 - \sum_{j=0}^{k-1} \frac{\exp(-(\lambda_2 - \lambda_1))(\lambda_2 - \lambda_1)^j}{j!} \right)
\end{aligned}
$$

See Janardan (1980) for an alternative derivation.

Table 1. *Descriptive Statistics*

	Mean	Std. Dev.	Min	Max
Number of doctor visits	2.058	2.810	0	25
No visit (yes/no)	0.343	0.475	0	1
Age	44.9	14.1	19	70
Years of schooling	12.1	2.50	7	18
Log net household income	10.32	0.643	6.68	13.46
Male (yes/no)	0.458	0.498	0	1
Disability (yes/no)	0.095	0.293	0	1
Current smoker (yes/no)	0.334	0.471	0	1
Body mass index	25.9	4.70	15.8	59.5

Source: Socio-Economic Panel (SOEP), version 26, doi:10.5684/soep.v26

Data for year 2006. N=2966.

Table 2. *Hurdle Poisson Models for Number of Doctor Visits*

	Poisson	Fixed Hurdle		Dynamic Hurdle	
	λ	λ_1	λ_2	λ_1	λ_2
Constant	0.451	-0.093	1.063	-0.088	1.125
	(0.409)	(0.457)	(0.264)	(0.448)	(0.385)
Age	0.003	-0.039	0.021	-0.038	0.038
	(0.011)	(0.012)	(0.007)	(0.012)	(0.010)
Age squared$\times 10^{-2}$	0.005	0.059	-0.022	0.056	-0.043
	(0.013)	(0.013)	(0.008)	(0.013)	(0.011)
Years of schooling	0.005	0.021	-0.007	0.022	-0.016
	(0.010)	(0.010)	(0.006)	(0.010)	(0.009)
Log net household income	-0.039	0.024	-0.068	0.024	-0.077
	(0.037)	(0.041)	(0.023)	(0.040)	(0.034)
Male (yes/no)	-0.272	-0.304	-0.142	-0.267	-0.137
	(0.050)	(0.051)	(0.029)	(0.049)	(0.042)
Disability (yes/no)	0.825	0.740	0.589	0.556	0.643
	(0.066)	(0.087)	(0.035)	(0.081)	(0.053)
Current smoker (yes/no)	-0.087	-0.161	-0.009	-0.129	-0.012
	(0.052)	(0.054)	(0.032)	(0.054)	(0.046)
Body mass index	0.015	0.010	0.012	0.006	0.017
	(0.005)	(0.006)	(0.003)	(0.005)	(0.004)
Log likelihood	-6754.5	-6153.1		-5961.5	
Number of observations	2966	2966		2966	

Source: Socio-Economic Panel (SOEP), version 26, doi:10.5684/soep.v26

Standard errors in parentheses.