Zurich Open Repository and Archive
University of Zurich
Main Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2017

A dynamic hurdle model for zero-inflated count data

Baetschmann, Gregori ; Winkelmann, Rainer

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.

DOI: https://doi.org/10.1080/03610926.2016.1146766

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-126311
Journal Article
Accepted Version

Originally published at:
Baetschmann, Gregori; Winkelmann, Rainer (2017). A dynamic hurdle model for zero-inflated count data. Communications in Statistics. Theory and Methods, 46(14):7174-7187.
DOI: https://doi.org/10.1080/03610926.2016.1146766

ACCEPTED MANUSCRIPT

A Dynamic Hurdle Model for Zero-Inflated Count Data*

Gregori Baetschmann
University of Berne

Rainer Winkelmann
University of Zurich

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, 8032 Zurich, Switzerland; email: rainer.winkelmann@econ.uzh.ch. We are grateful to two anonymous referees as well as Robert Jung, Johannes Kunz, Joao Santos Silva, and Kevin Staub for valuable comments, to Daniel Auer for very able research assistance, and to the German Institute for Economic Research (DIW Berlin) for providing access to the German Socio-Economic Panel data.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

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 extra zeros. The negative binomial model predicts a higher proportion of zeros than does the Poisson model, for a given mean. In many applications, however, this is still insufficient to account for the full amount of zeros. Moreover, the negative binomial model postulates that zeros and positive outcomes come from the same underlying process, whereas researchers have become increasingly interested in models, where this is not the case.

Existing zero-inflation approaches in the literature have focused on mixture models and two-part models. In a finite mixture model there are two types of observations 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). The hurdle model combines a binary model for the 0/1+ decision with a truncated-at-zero count data model for positive outcomes (Mullahy 1986). Both approaches include separate parameters, or regression coefficients, for the binary and count parts, respectively. By construction, they are designed to fit a pattern of excess zeros but not others, where higher counts are over- or underrepresented as well. These models have been regularly used in applied work, with 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 regression model for zero-inflated count data. Starting point is, as in the above approaches, a cross section of counts, i.e., the number of events in a fixed

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

time interval, where the timing of events is unobserved. As is well known, the Poisson distribution for the number of events arises if times between events are independently and identically exponentially distributed with constant hazard rate λ. The negative binomial models can be obtained by mixing the λ rate with a gamma distributed multiplicative heterogeneity term.

The modification we propose is to allow the hazard rate for the time to first event to differ from the hazard rate from the first to the second, second to the third event and so forth. 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). In the taxonomy of Heckman and Borjas (1980), our model allows for occurrence dependence but not duration dependence. Occurrence dependence means that the mere occurrence of an event alters the probability of future events. Duration dependence means that the time that has elapsed since the last event changes the probability of a further event (i.e. a departure from the Poisson process constant hazard rate assumption).

Very general patterns of occurence dependence are conceivable. We consider here the simplest case of a one-time change in the hazard rate, in the wake of the first event. The main reason is that we thereby obtain a model most similar and comparable to those developed in the aforementioned literature. Indeed, the binary 0/1+ model is identical to that implied by a hurdle-at-zero count model. Zero-inflation arise if the rate increases after the first event. In contrast to the standard 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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

second state. As a consequence, the probability of a zero and the distribution of positive outcomes are not independent. An extended version of the model also accounts for unobserved heterogeneity, where the rate of the process itself follows a gamma distribution.

Our “dynamic hurdle” approach has a number of useful properties. It nests the standard Poisson or negative binomial models so that testing for the absence of occurrence dependence is straightforward. The parameters are easy to interpret, as they indicate proportional shifts in the hazard rates, of the time to first event and the time span 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 briefly present the standard models for count dependent variables, including hurdle and zero-inflated count models. 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.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

2 Modeling zero-inflated count data

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

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

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 distribution 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 proportion of zeros. Similarly, extra zeros can be generated from a model where the time between events has a distribution with negative duration dependence (Winkelmann, 1995).

In many applications, extra zeros (relative to the Poisson model) as generated by the negative binomial model 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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

increase it. Moreover, one often has a substantive (structural) interest to treat the zero-generating process separately from the process for strictly positive outcomes, which requires different sets of parameters. This is related to the distinction between Tobit I and Tobit II models for non-negative continuous dependent variables with positive probability mass at zero.

For count data, there are two standard ways of accounting for extra zeros in Poisson or negative binomial models. In the fixed-hurdle (FH) model (Mullahy, 1986)

$$\Pr(Y = k) = \begin{cases} \phi_1 & \text{for } k = 0 \\ (1 - \phi_1)\dfrac{f(k)}{1 - f(0)} & \text{for } k = 1, 2, 3, \ldots \end{cases} \tag{2}$$

In the zero-inflated (ZI) model (Mullahy, 1986, Lambert, 1992)

$$\Pr(Y = k) = \begin{cases} \phi_2 + (1 - \phi_2)f(0) & \text{for } k = 0 \\ (1 - \phi_2)f(k) & \text{for } k = 1, 2, 3, \ldots \end{cases} \tag{3}$$

In either case does $f(k)$ denote the probability function of a standard count data model, usually either the Poisson or the negative binomial distribution. $f(k)$, ϕ_1 and ϕ_2 can be made dependent on covariates. These can be identical in the two parts of the model or distinct, and correlation is possible as well, in principle, although rarely allowed for in practice. In the application below, $\phi_1 = \exp(-\lambda)$ is modeled as survivor function of the exponential distribution, while ϕ_2 is of logit form.

A key difference between the two models is the origin of zeros: while there is a single type of zeros under the FH assumption, there are two types in the ZI model, sometimes referred to as “strategic” versus “incidental” zeros. Which of the two assumptions is preferable depends on the

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

specific application. Sometimes, a decision can be based on a-priori reasoning, in other cases, one can consider the issue ex-post, after having estimated the two models, by comparing the goodness-of-fit.

The current paper extends the hurdle type approach and thus speaks to the kind of application, where the presence of two types of zeros is implausible. In such hurdle models, 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, $1 - \phi_1$ is known as the utilization probability, i.e. the probability of using services at least once. When $(1 - \phi_1) > f(0)$, the data are zero-inflated relative to the base distribution.

While the hurdle 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 event occurred 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 time effect constitutes the key difference to the fixed hurdle model: the lower the state 1 rate, the later the expected time of crossing and the less time is spent in the state 2 process.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

3 Dynamic hurdle count models

The purpose of this section is to derive a new model for the total number of events that have occured in the interval $(0, T]$. This quantity will be denoted as $Y(0, T)$. We do not observe the times at which events have taken place. Let t, $0 < t \leq T$, induce a partition of the time interval. It follows that $Y(0, T) = Y(0, t) + Y(t, T)$.

3.1 Timing of the first event

There is a fundamental relationship between the time of the first event, denoted as ϑ_1, and the total number of events between 0 and T, $Y(0, T)$. If a first event occurs at $\vartheta_1 = t$ and $Y(t, T) = k - 1$, then, for $t \leq T$ and therefore $k \geq 1$, $Y(0, T) = k$. It follows that

$$\begin{aligned} \Pr[Y(0,T) = k, \vartheta_1 = t] &= \Pr[Y(t,T) = k-1, \vartheta_1 = t] \\ &= \Pr[Y(t,T) = k-1] f_1(t) \qquad (4) \end{aligned}$$

where $f_1(t)$ is the density function of the time of the first event. In our set-up, the time of the first event is unobserved, and our model therefore focusses on the marginal probability function $\Pr[Y(0, T) = k]$, which can be obtained as

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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

The first term averages over the probabilities of $k-1$ events occuring in the interval (t, T) for all possible values of t, using the density function of ϑ_1 as weights. The second term says that the probability of no event is equivalent to the probability of the first event occuring later than T.

The dependence of the distribution of $Y(0, T)$ on the time of first event t is easiest seen by considering the expected value $\mathrm{E}(Y(0, T))$:

$$\begin{aligned}
\mathrm{E}(Y(0,T)) &= \sum_{k=1}^{\infty} k \int_0^T \Pr[Y(t,T)=k-1] f_1(t) dt \\
&= \int_0^T \sum_{k=0}^{\infty} (k+1) \Pr[Y(t,T)=k] f_1(t) dt \qquad (5) \\
&= \Pr(y>0) + \mathrm{E}_t[\mathrm{E}Y(t,T)]
\end{aligned}$$

The first equality defines the expectation of a count variable, a probability weighted sum over all possible outcomes, where the term for $k = 0$ can be ignored. The second equality changes the order of integration and summation as well as the summation index. The third equality follows since $\Pr(\vartheta_1 < T)$ is equal to the probability of at least one event.

Hence, the mean of a count variable is equal to the probability of a positive count plus the expected number of events occurring after the time of the first event, i.e., after crossing the dynamic hurdle. Since the $\mathrm{E}Y(t, T)$ term is a decreasing function of t, the overall mean decreases as the expected time of the first event increases. This is a completely general result obtained without imposing any particular structure on the stochastic process.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

3.2 Dynamic hurdle Poisson model

To obtain the dynamic hurdle Poisson model, we make the following assumptions. First, $f_1(t)$ is assumed to be exponentially distributed with rate λ_1. 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 $k \geq 1$,

$$\Pr[Y(0,T)=k] = \int_0^T \frac{\exp(-\lambda_2(T-t))[\lambda_2(T-t)]^{k-1}}{(k-1)!}\lambda_1 \exp(-\lambda_1 t)dt \tag{6}$$

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

$$f_{DHP}(k;\lambda_1,\lambda_2) = \frac{\lambda_1 \lambda_2^{k-1}\exp(-\lambda_1)}{(\lambda_2-\lambda_1)^k}\left[1-\sum_{j=0}^{k-1}\frac{\exp(-(\lambda_2-\lambda_1))(\lambda_2-\lambda_1)^j}{j!}\right] \tag{7}$$

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

3.3 Properties

The mean of the count variable depends on two terms, $\Pr(\vartheta_1 > 1)$ and $\mathrm{E}_t[\mathrm{E}Y(t,1)]$ (see (5)). For the dynamic hurdle Poisson model (7), the expected time spent in the first state has two components: either, the arrival time of the first event ϑ_1 is greater than 1. Then the time spent in the first state is 1, with probability $\Pr(\vartheta_1 > 1) = \exp(-\lambda_1)$. Or else, the arrival time of the first event ϑ_1 is smaller

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

than 1. In this case $t = \vartheta$ and

$$\mathrm{E}(t, t < 1; \lambda_1) = \int_0^1 \exp(-\lambda_1 t)\lambda_1 t \, dt = \lambda_1^{-1} - (1 + \lambda_1^{-1}) \exp(-\lambda_1)$$

It follows that

$$\begin{aligned} \mathrm{E}(t; \lambda_1) &= \exp(-\lambda_1) + \lambda_1^{-1} - (1 + \lambda_1^{-1}) \exp(-\lambda_1) \\ &= \lambda_1^{-1}(1 - \exp(-\lambda_1)) \qquad (8) \end{aligned}$$

and the expected time spent in the second state is therefore

$$\mathrm{E}(1 - t; \lambda_1) = 1 - \lambda_1^{-1}(1 - \exp(-\lambda_1)) \qquad (9)$$

We can use (9) to rewrite (5) as

$$\begin{aligned} \mathrm{E}_{DHP}(Y; \lambda_1, \lambda_2) &= \Pr(Y > 0; \lambda_1) + \lambda_2 \mathrm{E}(1 - t; \lambda_1) \\ &= [1 - \exp(-\lambda_1)] + \lambda_2[1 - \lambda_1^{-1}(1 - \exp(-\lambda_1))] \qquad (10) \\ &= \lambda_2 + (1 - \lambda_2/\lambda_1)[1 - \exp(-\lambda_1)] \end{aligned}$$

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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

3.4 Comparison to the fixed hurdle Poisson model

Equation (10) 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

$$\begin{aligned} \mathrm{E}_{FHP}(Y;\lambda_1,\lambda_2) &= \Pr(Y>0;\lambda_1)\mathrm{E}(Y|Y>0;\lambda_2) \\ &= [1-\exp(-\lambda_1)]\,\frac{\lambda_2}{1-\exp(-\lambda_2)} \end{aligned} \tag{11}$$

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,

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

$\mathrm{E}(Y_i|x_i)$, the conditional probability of a zero, $\Pr(Y_i = 0|x_i)$, or the conditional-on-positives mean, $\mathrm{E}(Y_i|Y_i > 0, x_i)$. 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_j a conformable parameter vector. This parameterization ensures positive rates and implies a semi-elasticity interpretation for β_j. Further it allows to treat exposure T_i, the length of the observation period, as a standard covariate. Incorporating exposure explicitly in the model becomes necessary if T_i varies between individuals and therefore cannot be normalized to one.

3.6 Unobserved heterogeneity

In empirical applications, the variation of Y is often higher than that implied by a Poisson model, even if λ is allowed to depend on covariates. In the Poisson model, one can account for this "overdispersion" by introducing an additional non-negative multiplicative error term u such that, conditional on λ and u, Y follows a Poisson distribution with parameter $\lambda \times u$. Suppose that u is independently gamma distributed with mean 1 and variance α. Then it is well known that the distribution of Y, conditional on λ but unconditional on u is negative binomial (NB) with mean λ and variance $\lambda(1 + \lambda\alpha)$. Similarly, there are NB variants of the zero-inflated and hurdle count data models.

The DHP model can be extended along the same lines. Let u again denote a gamma distributed individual effect and assume that T is normalized to 1. If u equally affects both rates of the DHP model, the conditional probability of observing a count k is $f_{DHP}(k; \lambda_1 u, \lambda_2 u)$. Taking expectations

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

over u gives the unconditional probability function:

$$f_{DHNB}(k;\lambda_1,\lambda_2,\alpha) = \int_0^\infty f_{DHP}(k;\lambda_1 u,\lambda_2 u)\text{Gamma}(u;\alpha)du$$

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

with $p = (\alpha+\lambda_1)/(\alpha+\lambda_2)$ and Gamma$(u;\alpha)$ denoting the gamma density function with mean 1 and variance α. If $\lambda_2 > \lambda_1$, the term in squared brackets equals the complementary cumulative distribution function of a NegBin distribution. The mean of the DH model with unobserved heterogeneity is given by

$$\begin{aligned} \text{E}_{DHNB}(Y|\lambda_1,\lambda_2,\alpha) &= \int_0^\infty \lambda_2 u + (1-\lambda_2 u/\lambda_1 u)(1-\exp(-\lambda_1 u))\text{Gamma}(u;\alpha)du \\ &= \lambda_2 + (1-\lambda_2/\lambda_1)(1-\text{NB}(0;\lambda_1,\alpha)) \\ &= \lambda_2\text{E}(1-t,\lambda_1,\alpha) + \Pr(Y \neq 0,\lambda_1,\alpha) \end{aligned}$$

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

Of course, other approaches for incorporating unobserved heterogeneity would be available. For example, one could allow for two, potentially correlated, heterogeneity distributions, one for each state. Here, we make the polar assumption that the correlation is one. Furthermore, different mixing distributions are conceivable, such as a lognormal, and semi-parametric approaches are possible as well, for instance based on discrete mass points. Our proposed specification has two advantages. First, it is relatively simple and leads to a closed form probability function. And

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

second, our assumption ensures that the dynamic hurdle model nests the standard negative binomial model.

3.7 Decomposing the mean effect

The FH model (see 11) 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:

$$\frac{\partial \mathrm{E}_{FH}(Y;\lambda_1(x),\lambda_2(x))}{\partial x} = \frac{\partial \Pr(Y>0;\lambda_1(x))}{\partial x}\mathrm{E}(Y|Y>0;\lambda_2(x)) + \frac{\partial \mathrm{E}(Y|Y>0;\lambda_2(x))}{\partial x}\Pr(Y>0;\lambda_1(x)) \tag{12}$$

It is useful to think of the extensive margin effect as a participation effect (e.g., 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*.

The dynamic hurdle model lends itself to a more detailed decomposition of marginal mean effects. Differentiating (10) with respect to x, the DH model implies the following decomposition of the partial derivative of the overall mean:

$$\frac{\partial \mathrm{E}_{DH}(Y;\lambda_1(x),\lambda_2(x))}{\partial x} = \frac{\partial \Pr(Y>0;\lambda_1(x))}{\partial x} + \lambda_2\frac{\partial \mathrm{E}(1-t;\lambda_1(x))}{\partial x} + \mathrm{E}(1-t;\lambda_1(x))\frac{\partial \lambda_2(x)}{\partial x} \tag{13}$$

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

ACCEPTED MANUSCRIPT

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 conditional-on-positives effect can now further be decomposed into a time effect and an intensity effect.

3.8 Estimation and testing

One can estimate β_1 and β_2 jointly by maximum likelihood and use standard ML properties to compute standard errors (or robust standard errors if one considers Quasi-ML estimation) and confidence intervals. In our application below, we use data from a random sample of close to 3000 persons, justifying the reliance on large sample approximations.

In empirical applications, the interest is often in testing for the presence of excess zeros. Under the null hypothesis of no additional zeros, $\lambda_1 = \lambda_2$ which requires that $\beta_1 = \beta_2$, and the DH Poisson model simplifies to a simple Poisson model. The likelihood ratio test statistic is chi-squared distributed with k degrees of freedon. A similar test is possible for the DHNB or FH models.

Since the DHP, FHP, and ZIP models, or their negative binomial variants, 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 DHP and FHP 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

ACCEPTED MANUSCRIPT

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 Kullback-Leibler distance (see Vuong, 1989, for additional detail). Alternatively, one can select the best model using an information criterion.

Finally, there is a possibility of an informal specification test of the DHP model. Define the binary event “positive count yes/no”. Under the assumption of the DHP 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 DHP model. An informal specification test can be based on a comparison of $\hat{\beta}_1$ in the full DHP model with that of a simple binary model (i.e., the first stage of the FHP model). Large differences speak against the DH specification.

4 Application: determinants of doctor visits in Germany

This section reports results from an application to the socio-economic determinants of the frequency of doctor visits in Germany. The data for the analysis were extracted from the Socio-Economic Panel (SOEP, see Wagner et. al, 2007). We use data from a single cross-section for 2006, since a number of questions on interesting health behaviors was included in that year but not in others. We limit the sample to those covered by statutory health insurance, aged between 18 and 70. The final estimation sample has 2966 observations. Most SOEP interviews are conducted throughout the first half of the year. As part of the survey, respondents are asked for the number of times they have visited during the prior three-months period. This count is our depen-

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

dent variable. As Figure 1 shows, the number of visits is distributed between 0 and 25 visits, and 34 percent of all persons in the sample did not visit a doctor during the previous quarter. The mean is 2.06 visits and the standard deviation 2.8. The constant-only Poisson model would predict $\exp(-2.06) = 12.7\%$ of zeros, and it is unlikely that observed heterogeneity alone can account for the much larger proportion of zeros observed in the data.

The right panel of Figure 1 shows boxplots for four continuous regressors used in the analysis, namely income, age, years of schooling and the body mass index (BMI). Income is the log of net annual household income. The median BMI is 25.2, just above the lower bound of overweight. To allow for a potentially non-linear effect of age on doctor visits, we use a second-order polynomial in age. In addition, we include the binary regressors male (45.8% yes), disabled (9.5% yes) and current smoker (33.4% yes). Thus, including a constant, the hazard rate in both parts of the model depends on 9 parameters each.

A total of eight models were fitted to the data: the model with single index, fixed hurdle, dynamic hurdle and zero-inflation, each in their Poisson and negative binomial varieties. For the fixed hurdle negative binomial model, we used the complementary log-log model for the hurdle step (i.e. a Poisson-type hurdle) to avoid the identification issue raised in Pohlmeier and Ulrich (1995). Table 1 shows the log likelihood values and the associated values of the Schwarz information criterion. For all four models, there is strong evidence for the presence of unobserved heterogeneity. Since the hypothesis $\alpha = 0$ (where α is the variance of u) is at the boundary of the parameter space, the likelihood ratio test (LRT) statistic has a non-standard distibution, with 0.5 probability mass at zero and a $0.5 \times \chi^2(1)$ distribution for positive outcomes (Chernoff, 1954). Thus, to test at the 5% level

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

of significance, one should use the 90th percentile of a chi-squared distribution as critical value.

Furthermore, the single index Poisson model is rejected against the FHP and DHP models, as is the negative binomial model against FHNB and DHNB (e.g., in the latter case, the LRT statistic is 110, compared to a 5% critical value from the $\chi^2(9)$ distribution of 16.92). Among the three double index models that allow for extra zeros and flexible extensive margin effects, the DHNB is the model with the highest log likelihood value (-5,537.8 as compared to -5,539.5 for the FHNB and - 5,561.1 for the ZINB models). Thus, any of the standard model selection criteria would pick the DHNB model, and the same is true when applying the Vuong test for overlapping models.

The full set of estimated coefficients for the DHP and DHNB models is provided in Table 2. Columns (1) and (2) show the results for the hazard rates λ_1 and λ_2 for the two states of the DHP model, and columns (3) and (4) the corresponding results for the DHNB. A positive β means that a unit increase in the associated regressor increases the baseline hazard rate of the state specific process (which is constant in the DHP model and decreasing in the DHNB model) by $[\exp(\beta) - 1] \times 100$ percent.

There are some interesting asymmetries between state 1 (λ_1) and state 2 (λ_2) effects in the DHNB model. For instance, income has no effect at the state 1 extensive margin, but a statistically significant negative effect on the number of subsequent doctor visits, 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 state 1 effect is negative, while there is no effect on λ_2. Individuals with disabilities have higher hazard rates in both

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

states, and thus an unambiguously higher predicted number of doctor visits than others. Women have more visits than men, ceteris paribus. The DHP model leads to qualitatively similar conclusions. Importantly, the average predicted state 1 hazard rate is below the state 2 hazard rate in both models, indicating the presence of zero inflation.

While within-model comparisons of state 1 and state 2 parameters are meaningful, a comparison of, say, λ_1^{DHP} and λ_1^{DHNB} is not (note that the DHNB coefficients tend to be smaller than the DHP ones). The reason is that the baseline hazard rates differ in the two specifications, and coefficients translate differently into marginal effects. Comparable effects for all estimated models (including DHP and DHNB) are plotted in Figure 2. For the continuous regressors *log income* and *BMI*, we show the estimated effects of a one-standard deviation increase above the mean; for the binary regressors *smoke* and *disability*, we show the estimated effects of a change from 0 to 1. All other regressors are kept constant at their sample averages. Mean effects and standard deviations are obtained from a parametric bootstrap, drawing from the asymptotic distribution of the maximum likelihood estimators. The top panel of Figure 1 gives the effect on the probability of a zero, while the bottom panel displays the conditional-on-positives effect (i.e., the effect on $\mathrm{E}(Y|Y > 0, x)$).

The general conclusion is that the estimated effects of the various hurdle models are often quite similar, as are those of the zero-inflated models, with some exceptions for the ZINB. In contrast to that, the effects of the simple Poisson and negative binomial models show some larger deviations. For example, the predicted conditional-on-positives income effect is much smaller in the Poisson and NB models than in the extensions that allow for zero-inflation. The same holds true for the effect of smoking at the extensive margin. Perhaps surprisingly, smoking increases the probability

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

of no visit in all models, with point estimates being as high as 5.3 percentage points in the DHNB model. A possible explanation is that smokers are less health conscious than others, reducing their investments in preventive health care measures that may require doctor visits as well.

Another observation emanating from Figure 2 is the varying precision of the estimates. First, error bounds for the models without unobserved heterogeneity tend to be smaller, reflecting the known fact that neglected heterogeneity causes a downward bias in standard errors. Second, the fixed and dynamic hurdle negative binomial hurdle effects are estimated with similar precision, so that there is no loss from estimating the dynamic hurdle model. The DHP and DHNB models have a slightly more complex likelihood function but they in return provide interpretations related to the underlying stochastic process, such as the expected duration spent in the second state. For example, the DHP model has an average predicted state 1 hazard rate of 1.1. This corresponds to an average duration of 35 days spent in state 2. If the hazard rate increases by 10 percent to 1.21, the predicted duration in state 2 increases to 38 days.

5 Concluding remarks

This article develops a new approach for the regression analysis of zero inflated and overdispersed count data. In our dynamic hurdle 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, which is endogenously determined. Regressors are allowed to differentially affect the two hazard rates before and after the

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

first event, and the effect on the overall count has three distinct channels: the probability of a zero, the expected duration of the second state, and the hazard rate in that second state.

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 somewhat better than two existing approaches for zero-inflated count data. The results are easy to interpret, since regression coefficients are semi-elasticities of two hazard rates, and they allow for a flexible modeling of extensive margin and conditional-on-positives effects.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

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, *Journal 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.

Chernoff, H. (1954), On the distribution of the likelihood ratio, *Annals of Mathematical Statistics* 25: 573-578.

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.

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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

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.

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.

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.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

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.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Appendix. 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 $\lambda_1 \neq \lambda_2$:

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

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

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

See Janardan (1980) for an alternative derivation.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 1. *Model comparisons*

	log likelihood	number of parameters	SIC
Poisson	-6,754.4	9	13,580.8
Negative binomial	-5,592.8	10	11,265.5
Fixed hurdle Poisson (FHP)	-6,153.0	18	12,450.0
Fixed hurdle negative binomial (FHNB)	-5,539.5	19	11,231.0
Zero-inflated Poisson (ZIP)	-6,150.8	18	12,445.6
Zero-inflated negative binomial (ZINB)	-5,561.1	19	11,274.2
Dynamic hurdle Poisson (DHP)	-5,961.5	18	12,066.9
Dynamic hurdle negative binomial (DHNB)	-5,537.8	19	11,227.6

Note: SIC computed as -2 × log likelihood + number of parameters $\times \ln N$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 2. *Dynamic Hurdle Models for Frequency of Doctor Visits*

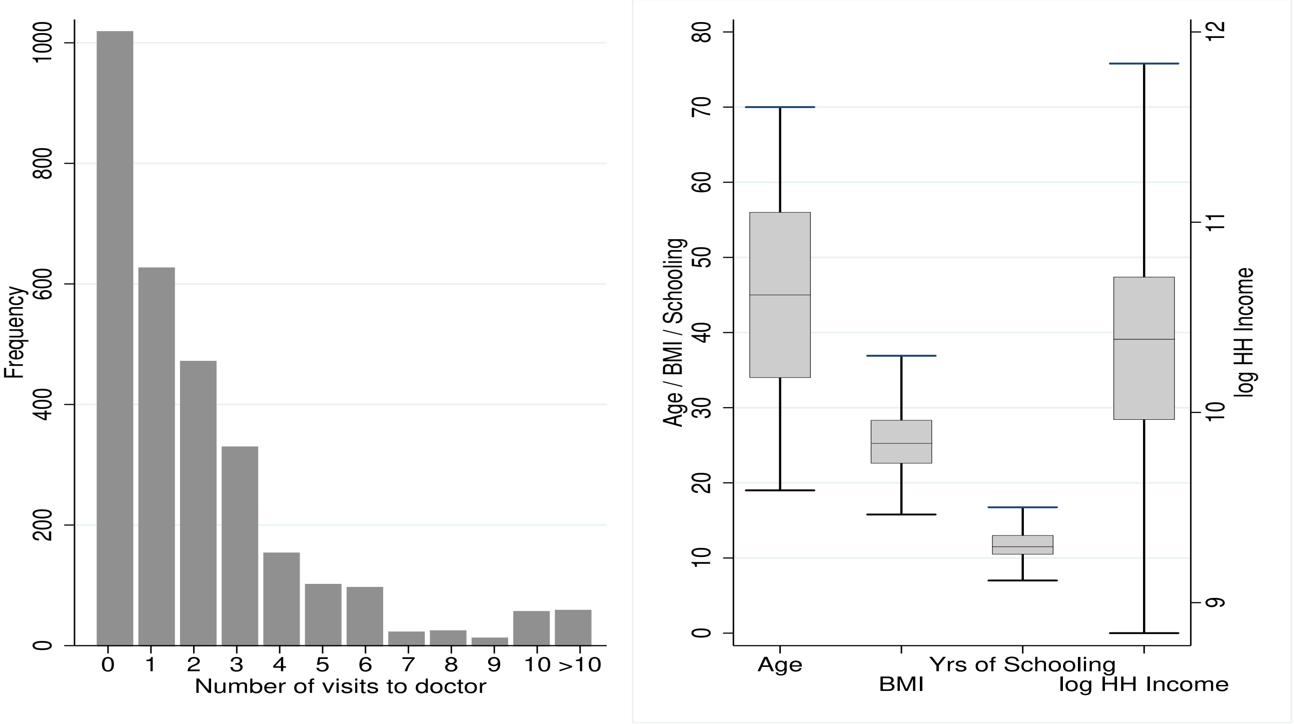
	no unobserved heterogeneity		with unobserved heterogeneity	
	λ_1^{DHP}	λ_2^{DHP}	λ_1^{DHNB}	λ_2^{DHNB}
Constant	0.436	0.604	-0.088	1.125
	(0.769)	(0.543)	(0.448)	(0.385)
Age	-0.081	0.024	-0.038	0.038
	(0.020)	(0.013)	(0.012)	(0.010)
Age squared $\times 10^{-2}$	0.118	-0.024	0.056	-0.043
	(0.023)	(0.014)	(0.013)	(0.011)
Years of schooling	0.037	-0.008	0.022	-0.016
	(0.018)	(0.012)	(0.010)	(0.009)
Log net household income	0.066	-0.077	0.024	-0.077
	(0.068)	(0.048)	(0.040)	(0.034)
Male (yes/no)	-0.520	-0.178	-0.267	-0.137
	(0.083)	(0.059)	(0.049)	(0.042)
Disability (yes/no)	1.601	0.717	0.556	0.643
	(0.231)	(0.081)	(0.081)	(0.053)
Current smoker (yes/no)	-0.239	-0.012	-0.129	-0.012
	(0.086)	(0.064)	(0.054)	(0.046)
Body mass index	0.016	0.017	0.006	0.017
	(0.010)	(0.006)	(0.005)	(0.004)
α				1.059
				(0.079)
Log Likelihood	-5,961.5		-5,537.9	
Number of observations	2,966		2,966	

Notes: DHP: dynamic hurdle Poisson (no unobserved heterogeneity); DHNB: dynamic hurdle negative binomial (with unobserved heterogeneity); Standard errors in parentheses. Data from 2006 wave of the German Socio-Economic Panel (version 26, doi:10.5684/soep.v26).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

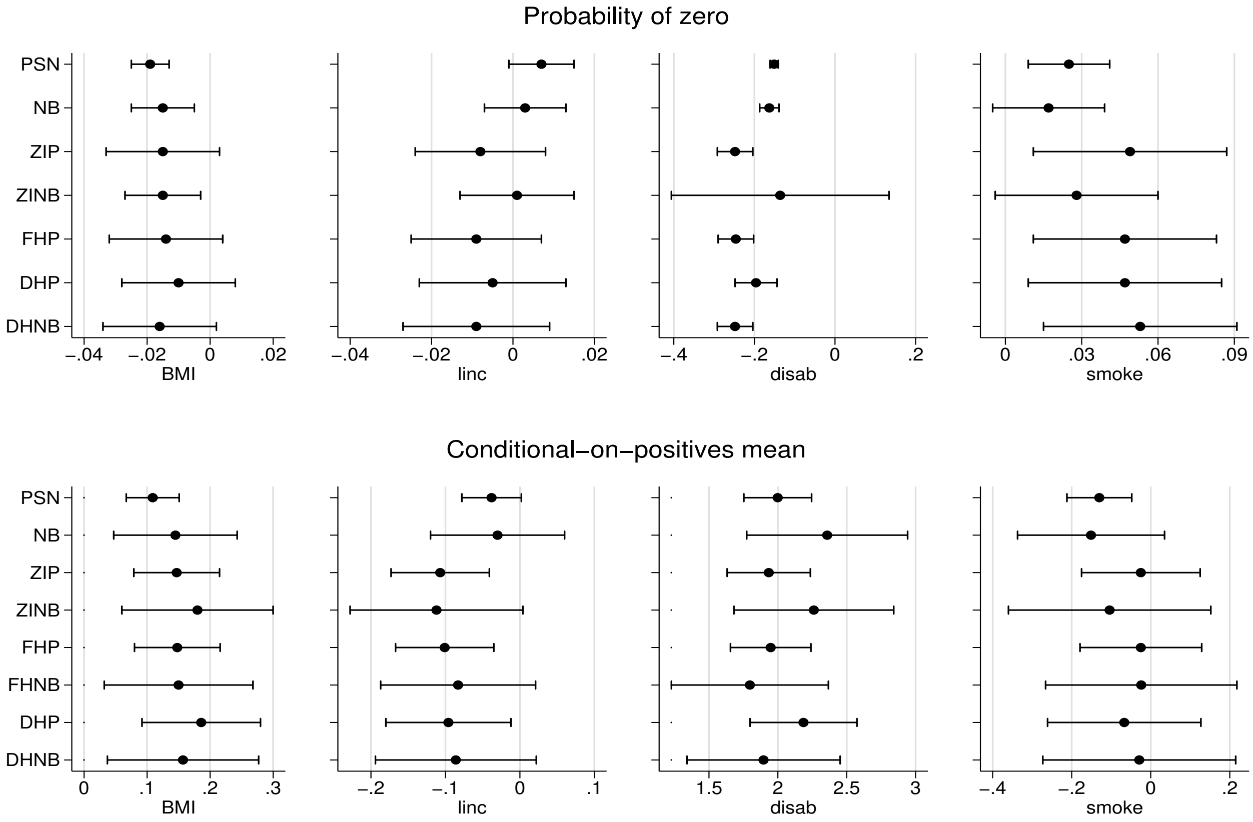
Figure 1: Summary statistics



ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Figure 2: Extensive margin and conditional-on-positives effects



ACCEPTED MANUSCRIPT