Modelling Panel Count Data With Excess Zeros: A Hurdle Approach

José M. R. Murteira[*]

CEMAPRE, and Faculdade de Economia, Universidade de Coimbra

Mário A. G. Augusto

Institute of Systems and Robotics, and Faculdade de Economia, Universidade de Coimbra

May, 2014

Abstract

This paper presents a hurdle-type regression model for panel count data with excess zeros and bounded support. Each time series of counts is modelled by making use of the binomial thinning, a conceptual device that facilitates the consideration of dependence between consecutive integers. Nonlinear least squares, quasi-maximum likelihood and maximum likelihood are suggested methods to estimate the resulting model, which can also be produced under a random effects approach, following suitable assumptions concerning the conditional distribution of unobserved individual heterogeneity. These assumptions yield a two-part panel data generalization of the beta-binomial model, estimable through user-friendly quasi-maximum likelihood. A Monte Carlo exercise illustrates the behaviour of the proposed estimators, both with and without unobserved heterogeneity, as well as the performance of some specification tests. Finally, the proposed approach is illustrated as a model of repayment behaviour, estimated with a panel data set on personal loans granted by a Portuguese financial institution.

JEL classification: C23, C25, G210

Key Words: panel count data; hurdle model; binomial thinning; quasi- maximum likelihood; beta-based mixture; loan repayment behaviour.

[*] Corresponding Author. Address: Av. Dias da Silva, 165, 3004-512 Coimbra, Portugal. E-mail: jmurt@fe.uc.pt. Fax: + 351 239790514.

1 Introduction

This paper presents two-part regression models for panel count data. Several features of the data generating process (DGP) require a modelling approach that differs from standard models available in the statistical and econometric literature on discrete data analysis. In particular, the proposed specification is intended for count data with bounded support and excess zeros relative to a one-part DGP, a kind of data that may occur, *e.g.*, in the areas of consumer credit or behavioural scoring, with respect to the repayment of loans through consecutive fixed installments. At each repayment date a debtor faces a known number of missed installments, which, on the one hand, depends on his previous repayment decisions and, on the other hand, is bounded by the age of the loan. At the same time, given that most clients usually repay their debts on time, the number of missed installments at each date is expected to display more zero values than would be the case if this variable were to follow a single-part distribution.

Throughout the text a sample of observations on independent time series of counts is supposed to be available to the researcher. For each time series, the possibility of excess zeros is dealt with through a hurdle-type specification (Cragg, 1971, Mullahy, 1986), which allows a differentiated treatment of null and positive values of the responses. Meanwhile, the bounded nature of the responses suggests that a model for limited dependent variables, such as the binomial distribution, is clearly more appropriate than such customary models as, *e.g.*, the Poisson or negative binomial.

The literature proposes several ways to model time series dependence among discrete variables: surveys can be found, among others, in McKenzie (2003) and Cameron and Trivedi (2013, Ch. 7). Integer-valued autoregressive, moving average (INARMA) models constitute a prominent, well established example, including INAR models as a special case. The latter specify the realized value of the variable of interest at period t, y_t, as the sum of an integer function of past outcomes and the realization of an independent integer random variable. These models have the same serial correlation structure as linear ARMA models for continuous data, a clearly attractive feature. Different choices for the distribution of the innovation term (*e.g.* Poisson) lead to different marginal probability functions (p.f.) for y_t. Within the pure time series case, the Poisson INAR(1) was first proposed by McKenzie (1985) and further discussed, along with other INARMA models, by Al-Osh and Alzaid (1987) and McKenzie (1988). The extension of these models to the regression case was initiated by Brännäs (1995), who proposed the specification of INAR parameters as functions of covariates. Meanwhile, one alternative route to allow for dependence among discrete variables consists on assuming that these variables share common, often time-dependent, unobservable features. Such is the case of Hidden Markov Models, described and illustrated in detail in MacDonald and Zucchini (1997).

In this text the dependence structure within each time series is addressed through the use of the "binomial thinning" operator (Steutel and VanHarn, 1979). As described below, this operator provides a flexible way to model the dependence between counts, with the observed integer for one period defining the upper bound of the support of the count variable in the following period. Binomial thinning constitutes the most popular form of thinning, a general probabilistic operation that can be applied to random counts.[(1)] The main idea is that the count represents the size of some population, randomly shrinked (or 'thinned') through the thinning operation. As the size of the thinned population is still integer-valued, the application of thinning always yields integer values. Thus (contrarily to conventional ARMA processes), all elements of the time sequence are integer-valued.

The remainder of the paper is organized as follows. Section 2 presents the general approach and illustrates its use as a model of loan-repayment behaviour. Section 3 proposes estimation of the model through pooled nonlinear least squares (NLS), pooled quasi-maximum likelihood (QML) and maximum likelihood (ML), and discusses procedures to test its specification. The performance of the proposed estimators is assessed in Section 4 on the basis of simulated data. Section 5 illustrates the application of the proposed model to a data set on personal loans granted by a Portuguese financial institution. Section 6 concludes and suggests future research and extensions. A final Appendix presents algebraic derivations and proofs of some of the formulas and results used in the text.

2 General Model

Assume the availability of a sample of observations on n independent time series of counts, $\boldsymbol{y}_i \equiv (y_{i1}, \ldots, y_{iT_i})'$, $i = 1, ..., n$, each sequence with $t = 1, \ldots, T_i$ terms. In what follows a model for each individual sequence of counts is presented, enabling the specification of a time series count data model. The first part of the Section describes a pure time series

[1] A survey of thinning operations and their applications can be found in Weiss (2008).

framework, appropriate to accomodate the introduction of explanatory variables and the construction of a regression model, as discussed in Section 2.2. Estimation and inference issues are discussed in Section 3.

2.1 Time Series Framework

For each individual i, the proposed specification starts with the structural assumption on the triplet $(y_{it}, y_{i,t-1}, d_{it})$

$$y_{it} = \begin{cases} y_{i,t-1} + 1, & d_{it} = 1 \\ 0, & d_{it} = 0, y_{i,t-1} = 0 \\ p_1 \circ y_{i,t-1}, & d_{it} = 0, y_{i,t-1} > 0 \end{cases}, \quad y_{it} \in \{0, 1, ..., t\}, t = 1, ..., T_i. \tag{1}$$

In this expression $y_{i0} \equiv 0$, $d_{it} \in \{0, 1\}$ denotes a Bernoulli random variable with $\Pr(d_{it} = 1) = p$, and the parameter p_1 is such that $0 \leq p_1 \leq 1$. The symbol "$\circ$" denotes the binomial thinning operator, defined as follows. If z denotes a positive integer, then $p_1 \circ z \equiv \sum_{j=1}^{z} b_j(p_1)$, where $\{b_j(p_1), j = 1, ..., z\}$ denotes a set of i.i.d. Bernoulli random variables, with $\Pr(b_j(p_1) = 1) = p_1$. Thus, given z, $p_1 \circ z$ represents a binomial random variable, the number of successes in z independent trials in each of which the probability of success is p_1.

Consequently, for each individual (for simplicity, the index i is omitted in the remainder of the present Section) the support of y_t, given y_{t-1}, is $\{0, 1, ..., y_{t-1} + 1\}$: if $d_t = 1$, then $y_t = y_{t-1} + 1$; on the contrary, if $d_t = 0$, then, given y_{t-1}, y_t is either zero (if $y_{t-1} = 0$) or follows a binomial p.f. with parameters y_{t-1} and p_1. The probabilistic model for $y_t|y_{t-1}$ is thus specified as a two-part, or hurdle, model: the first part is a binary outcome model and the second part is a count model with bounded support. Formally,

$$\Pr(y_t|y_{t-1}) = E_d(\Pr(y_t|y_{t-1}, d_t)) = \begin{cases} p, & y_t = y_{t-1} + 1, \\ (1-p)\binom{y_{t-1}}{y_t} p_1^{y_t}(1-p_1)^{y_{t-1}-y_t}, & y_t \in \{0, 1, ..., y_{t-1}\} \end{cases} \tag{2}$$

Unless $p = 0$ (corresponding to the trivial case $y_t \equiv 0$, $\forall t$), the sequence $\boldsymbol{y} \equiv (y_1, \ldots, y_T)$ is not stationary. To see this start by considering the sequence of conditional moment generating functions (mgf), of y_t, given y_{t-1}, which, under the model's assumptions and the definition of the binomial thinning operator can be written as

$$M_{t|t-1}(s) \equiv E(\exp(sy_t)|y_{t-1}) = E_d(E(\exp(sy_t)|y_{t-1}, d_t)) =$$
$$pe^{s(y_{t-1}+1)} + (1-p)(1-p_1+p_1e^s)^{y_{t-1}}. \tag{3}$$

By evaluating derivatives of $M_{t|t-1}(s)$ at $s=0$, one can produce conditional moments of any order, $E\left(y_t^k|y_{t-1}\right)$, $k \in \mathcal{N}$. The first two conditional moments are given by

$$E(y_t|y_{t-1}) = p + ry_{t-1}, \tag{4}$$
$$E\left(y_t^2|y_{t-1}\right) = p + (2p + p_1(1-p_1)(1-p))y_{t-1} + \left(p + p_1^2(1-p)\right)y_{t-1}^2, \tag{5}$$
$$t = 1, ..., T_i,$$

where $r \equiv p + p_1(1-p)$. From these results one can obtain, *e.g.*, the conditional variance $V(y_t|y_{t-1}) = E(y_t^2|y_{t-1}) - E^2(y_t|y_{t-1})$.

The unconditional first moment of y_t can then be obtained by successively applying the law of iterated expectations; using the initial condition $y_0 \equiv 0$,

$$E(y_t) = E(E(y_t|y_{t-1})) = p + rE(y_{t-1}) =$$
$$p + rE(E(y_{t-1}|y_{t-2})) = p(1+r) + r^2E(y_{t-2}) =$$
$$\ldots = p\left(1 + r + \ldots + r^{t-1}\right) + r^tE(y_0)$$
$$= p\left(1 + r + r^2 + \cdots + r^{t-1}\right). \tag{6}$$

If p and p_1 are both less than 1 (so $r < 1$), this equals $E(y_t) = p(1-r^t)/(1-r)$. Aside from the trivial case $p = 0$, $E(y_t)$ varies with t, so the sequence $\boldsymbol{y}$ is not stationary.

The covariance function can be expressed as

$$COV(y_t, y_{t-k}) = r^kV(y_{t-k}), \tag{7}$$

a similar result to the expression for the covariance function in linear AR models for continuous variables (see the Appendix for details). In this expression the variance can be obtained from $V(y_t) = E(y_t^2) - E(y_t)^2$ with $E(y_t^2)$ expressed as the solution to the difference equation (obtained from (5))

$$E\left(y_t^2\right) = E\left(E\left(y_t^2|y_{t-1}\right)\right) = \tag{8}$$
$$p + (p + r(1-p_1))E(y_{t-1}) + (p + rp_1)E\left(y_{t-1}^2\right),$$

with $E\left(y_{t-1}\right)$ given in (6) and initial condition $E\left(y_1^2\right) = p$. It is then evident that the expression for $V\left(y_t\right)$ also depends on t.[2] Consequently, the autocorrelation function

$$CORR\left(y_t, y_{t-k}\right) = r^k \sqrt{V\left(y_{t-k}\right)/V\left(y_t\right)}, \tag{9}$$

depends not only on lag (k) but also on t.

For each individual the present model can be viewed as a nonstationary version of the DAR(1) model proposed by Jacobs and Lewis (1978). The latter can be expressed as

$$y_t = d_t y_{t-1} + \left(1 - d_t\right) z_t,$$

where d_t are i.i.d. binary and z_t are i.i.d. with a given distribution. If y_0 is also sampled from this distribution, then the model defines a stationary process with that same marginal distribution. Model (1), however, differs from this approach because not only stationarity is absent (due to the term $y_{t-1} + 1$) but also the variable $p_1 \circ y_{t-1}$, replacing z_t, is not i.i.d..

The proposed model can also be cast within the general framework of a first-order Markov chain. For the time sequence of counts one can formally write the following recurrence formula for $\Pr\left(y_t\right)$:

$$\Pr\left(y_t = y\right) = \sum_{j=\max\{1,y\}}^{t} \Pr\left(y_t = y | y_{t-1} = j-1\right) \Pr\left(y_{t-1} = j-1\right),$$
$$0 \leq y \leq t, \quad 1 \leq t \leq T,$$

with $\Pr\left(y_0 = 0\right) = 1$ and transition probabilities expressed in (2),

$$\Pr\left(y_t = y | y_{t-1} = j-1\right) = \begin{cases} p, & y = j, \\ (1-p)\binom{j-1}{y} p_1^y \left(1-p_1\right)^{j-1-y}, & y \in \{0, 1, \ldots, j-1\}. \end{cases}$$

2.2 Regression Model

A regression model can be produced by introducing explanatory variables in the probabilities p and/or p_1, following the usual practice for discrete choice models. These probabilities can be specified as, *e.g.*, logit, probit or any other convenient model, based on some chosen c.d.f..

[2] Trivially, if $p = 1$ ($\Leftrightarrow r = 1$), then $E\left(y_t\right) = t$ and if $p = 0$, then $E\left(y_t\right) = 0$. In both cases y_t is degenerate so $V\left(y_t\right) = 0$.

The expectation of y_t can be taken either conditionally on $\boldsymbol{y}_{(t-1)} \equiv (y_1, \ldots, y_{t-1})$ (name this the 'autoregressive model' or 'autoregressive approach'), or marginally to $\boldsymbol{y}_{(t-1)}$ (name this the 'marginal model' or 'marginal approach'). Under an 'autoregressive' approach the conditional mean of interest will be $E\left(y_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$, where $\boldsymbol{x}_t$ denotes the explanatory variables in the conditional expectation of y_t apart from lags of the dependent variable (that is, the covariates in p and p_1, other than lagged responses). Under a 'marginal' approach the conditional expectation of interest is $E\left(y_t|\boldsymbol{x}_{(t)}\right)$, where $\boldsymbol{x}_{(t)} \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)$.

If all covariates in p and p_1 are time-invariant (only involving covariates of the type $\boldsymbol{x}_t \equiv \boldsymbol{x}$, $\forall t$), then the above formulae for the moments of the response variables (*e.g.*, (6), (7) and (4)) remain substantively unchanged – the only difference being that the probabilities p and p_1 are now time-invariant functions of covariates. If, however, p and/or p_1 involve time-varying covariates, then the former expressions have to be altered accordingly.

Write the first conditional moment of y_t given its own lags and $\boldsymbol{x}_t$ as

$$E\left(y_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right) = p_t + r_t y_{t-1}, \tag{10}$$

where $p_t \equiv p\left(\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$, $r_t \equiv r\left(\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right) \equiv p_t + p_{1t}(1-p_t)$ and $p_{1t} \equiv p_1\left(\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$. If p and p_1 do not involve lags of the dependent variable, the conditional expectation of y_t, given $\boldsymbol{x}_{(t)}$, is given by

$$E\left(y_t|\boldsymbol{x}_{(t)}\right) = \begin{cases} p(x_1) & t = 1, \\ p_t + \sum_{j=1}^{t-1} p_j \left(\prod_{k=j+1}^{t} r_k\right), & t \geq 2, \end{cases} \tag{11}$$

which, as expected, reduces to (6) when only time-invariant covariates are used in p and p_1 (see the Appendix). If, on the other hand, lags of the dependent variable are introduced in these probabilities, the expression of $E\left(y_t|\boldsymbol{x}_{(t)}\right)$ becomes more complex because then it corresponds to the marginal expectation of (10) with respect to $\boldsymbol{y}_{(t-1)}$, given remaining covariates.

At any given period t, marginal effects can be obtained from the derivatives of the mean of y_t (for continuous covariates) or its differences (for discrete explanatory variables). The particular expression of the mean to use in each case hinges on: (*i*) the type of covariates included in p and p_1 (namely, time-varying or time-invariant); (*ii*) whether the conditional or marginal expectation is used. It seems clear that a marginal approach can become very burdensome when there are time-varying covariates in p and/or p_1. Therefore, in this latter case the ensuing exposition only considers marginal effects on the conditional expectation $E\left(y_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$.

Consider, firstly, marginal effects on $E\left(y_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$ – expression (10). The marginal effect of a unit change of a continuous covariate (w, say) can be written as $\nabla_w p_t + \nabla_w r_t y_{t-1}$, with $\nabla_w r_t = (\nabla_w p_t)(1 - p_{1t}) + (\nabla_w p_{1t})(1 - p_t)$. If p and p_1 are both monotonically increasing (decreasing) functions of w – so $\nabla_w p_t$ and $\nabla_w r_t$ are both positive (negative) – this marginal effect is positive (negative). For a discrete covariate w, the marginal effect can be expressed as $\Delta_w p_t + \Delta_w r_t y_{t-1}$, with (using obvious notation) $\Delta_w p_t \equiv p_t(w+1) - p_t(w)$ and $\Delta_w p_{1t} \equiv p_{1t}(w+1) - p_{1t}(w)$, so that

$$\Delta_w r_t \equiv r_t(w+1) - r_t(w) = (1 - p_{1t} - \Delta_w p_{1t})\,\Delta_w p_t + (1 - p_t)\,\Delta_w p_{1t}.$$

Now, p and p_1 are both probabilities so, usually, $0 < 1 - p < 1$ and $0 < p_1 + \Delta_w p_1 < 1$ (omit t for now). If p and p_1 are both monotonically increasing in w ($\Delta_w p > 0$ and $\Delta_w p_1 > 0$), then $\Delta_w r > 0$; if p and p_1 are both decreasing in w ($\Delta_w p < 0$ and $\Delta_w p_1 < 0$), then $\Delta_w r < 0$. Then, from (10), if p_t and p_{1t} are strictly increasing (decreasing) functions of w, then a unit increase in w has a positive (negative) impact on $E(y_t|\boldsymbol{x}_t, y_{t-1})$.

Secondly, consider marginal effects on the marginal expectation, $E\left(y_t|\boldsymbol{x}_{(t)}\right)$ (as previously mentioned, only the case with time-invariant covariates is discussed, $\boldsymbol{x}_t = \boldsymbol{x}, \forall t$ – see (6)). For a continuous covariate w,

$$\nabla_w E(y_t|\boldsymbol{x}) = \begin{cases} \nabla_w p, & t = 1, \\ (\nabla_w p)\left(1 + r + \cdots + r^{t-1}\right) + p\left(1 + 2r + \cdots + (t-1)\,r^{t-2}\right)(\nabla_w r)\,, & t \geq 2. \end{cases} \tag{12}$$

Equivalently, with $r < 1$,

$$\nabla_w E(y_t|\boldsymbol{x}) = \begin{cases} \nabla_w p, & t = 1, \\ (\nabla_w p)\dfrac{1 - r^t}{1 - r} + p\dfrac{1 - tr^{t-1} + (t-1)\,r^t}{(1-r)^2}(\nabla_w r)\,, & t \geq 2. \end{cases}$$

If p and p_1 are both monotonically increasing (decreasing) functions of w – so $\nabla_w p$ and $\nabla_w r$ are both positive (negative) – the previous expressions are evidently positive (negative). A similar conclusion can be drawn with respect to the efect of a unit increase of a discrete covariate on $E(y_t|\boldsymbol{x})$ – check (6) and recall the above discussion on marginal effects of changes in a discrete covariate on the conditional mean.

2.3 Application: Modelling Loan Repayment Behaviour

If one thinks of individuals as borrowers repaying their loans through fixed periodic installments, with y_{it} denoting the number of missed installments by individual i at the end of period t, then the probabilistic model for $y_t|y_{t-1}$ can be seen as a model of repayment behaviour.[3] To this effect, let the binary variable d_t be defined as zero if the client decides to pay at period t (that is, 'success' refers to a *non*-payment decision, with probability denoted by p). Once the client chooses to pay ($d_t = 0$) he then decides how many installments to pay, ranging from just one ($p_1 \circ y_{t-1} = y_{t-1}$), to all the installments missed that far plus the one for the present period ($p_1 \circ y_{t-1} = 0$). That is, under the present framework each borrower is supposed to make a twofold decision at every repayment date: firstly, he decides whether to pay or not altogether at that date; then, if he chooses to pay and more than one installment are due, he decides how much (how many installments) to pay.

This example illustrates the possible meaning of several limiting (though some unlikely) cases regarding different values of p and p_1. Firstly, $p = 0$ (d_t degenerate at zero, $\forall t$) means the borrower always decides to pay, one installment at least. From $y_0 \equiv 0$ it follows $y_1 \equiv d_1 = 0$; given that $y_t = p_1 \circ y_{t-1}$ and the individual always pays, then $y_{t-1} = 0 = y_t$, $\forall t$, so p_1 is not identified. On the other hand, $p = 1$ means that the client always decides not to pay: then, $d_t = 1$, $\forall t$, so $y_t = y_{t-1} + 1 = t$ and p_1 is again not identified.

Next, consider the case $p_1 = 0$. Now, once the client decides to pay ($d_t = 0$) he pays all the missed payments at that time (all b_j's are degenerate at zero), in addition to that period installment. That is, at each repayment date a borrower either pays all the installments he missed that far, plus the one for that period, or he pays none. On the other hand, $p_1 = 1$ means that when the client decides to pay ($d_t = 0$) he only pays one installment (all b_j's degenerate at one). That is, with $p_1 = 1$ the client's balance of missed payments never decreases – at best it remains at the level it rose to, the last time the client chose not to pay.

Naturally, some of the above limiting cases may be unlikely – namely the case with $p = 1$. Also, even for those (presumably many) borrowers who regularly meet their obligations,

[3] A survey on credit and behavioural scoring techniques can be found, *e.g.*, in Thomas, Edelman and Crook (2002).

p can be positive due to the possible occurrence of random unexpected (and undesirable) events, affecting their financial capability and hindering their plans to pay one installment after the other during the life of the contract. Nonetheless, a well known advantage of the hurdle specification is that it enables a differentiated treatment of zero and positive values of the dependent variable. In the present example, with most individuals expected to regularly meet their financial obligations, this feature seems clearly attractive.

It may be more appropriate to consider p and p_1 individual-specific, rather than constant across different individuals. Also, a pure time series model may not lead to very useful conclusions – for instance, credit or behavioural scoring purposes may require the classification of debtors on the basis of contracts and customers' characteristics. In order to account for this individual heterogeneity one obvious possibility is to introduce regressors in the model (*e.g.*, in p and/or p_1). In addition, further dynamics can be allowed for by considering lagged dependent variables in the specification of p and or p_1 (for instance, the probability of nonpayment may depend negatively on the number of previously missed installments).

The assumption of independence about the joint p.f. of $(d_t, y_{t-1}, b_1, ..., b_k)$, $k = y_{t-1}$, may also be inappropriate or unrealistic. On the other hand, analysis and estimation of the model is considerably eased with such an assumption. A simplifying way out is to assume that d_t and y_{t-1} are independent conditionally on p. Also, b_j, $j = 1, ..., t-1$ and y_{t-1} can be assumed conditionally independent, given p_1. Then, the independence assumption of these variables can be relaxed with the introduction of regressors in the model, in p and/or p_1. If, for instance, p and p_1 are specified as logits with time-invariant regressors, one will have

$$p_i \equiv \left(1 + \exp\left(-\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{-1}, \qquad p_{1i} \equiv \left(1 + \exp\left(-\boldsymbol{z}_i'\boldsymbol{\gamma}\right)\right)^{-1}, \tag{13}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ denote vectors of parameters conformable to, respectively, covariates' vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$.

3 Estimation and Inference

This Section discusses estimation and testing of the model presented above, on the basis of a sample of n independent time series of response variables and covariates $\{(y_{it}, \boldsymbol{x}_{it}),$

$t = 1, \ldots, T_i,\ i = 1, \ldots, n\}$. Throughout the Section the case with n large relative to $\max\{T_i, i = 1, \ldots, n\}$ is considered, as the asymptotics hold with bounded T_i and $n \to \infty$.

Al-Osh and Alzaid (1987), Jin-Guan and Yuan (1991) and Ronning and Jung (1992) discuss the use of least squares and ML methods in the context of the stationary Poisson INAR. Brännäs (1994) introduces GMM and extends its use to a panel data generalized Poisson INAR(1). Brännas (1995) proposes a Poisson INAR(1) regression model and studies its estimation through conditional least squares and GMM. More recent accounts of estimation methods for panel count data can be found in Jung, Kukuk and Liesenfeld (2006), Windmeijer (2006) and Sun and Zhao (2013 – in the area of Bio- and Health Statistics).

As mentioned, the conditional expectation of y_{it} can be taken either conditionally on $\boldsymbol{y}_{i(t-1)} \equiv (y_{i1}, \ldots, y_{i,t-1})$ ('autoregressive model'), or marginally to $\boldsymbol{y}_{i(t-1)}$ ('marginal model'). If all covariates are time-invariant, adoption of either an autoregressive or a marginal approach (using, respectively, (4) or (6)) seems fairly atraightforward. On the other hand, with time-varying covariates the marginal approach can become very burdensome so, in this case, the autoregressive model (10) seems clearly more attractive from a practical standpoint.

3.1 Pooled Nonlinear Least Squares

Among other possibilities, the proposed model can be estimated by pooled NLS. This method minimizes the criterion $\sum_{i=1}^{n} \sum_{t=1}^{T_i} (y_{it} - \mu_{it})^2$ with respect to the parameters in p and p_1, where μ_{it} denotes the appropriate ('autoregressive' or 'marginal') expression of the conditional expectation of y_{it}. This objective function is the one associated with pooled least squares estimation of the model

$$y_{it} = \mu_{it} + u_{it}, \quad t = 1, ..., T_i, \quad i = 1, ..., n,$$

where, by definition, the error term, u_{it}, has null conditional mean. Consequently, under standard regularity assumptions, for bounded T_i, $i = 1, ..., n$, and $n \to \infty$ the pooled NLS estimator is consistent and $\sqrt{n}$-asymptotically normal.

By construction the errors of this model are heteroskedastic and uncorrelated (across both i and t): with regard to error variance note that both the conditional and marginal variances of the dependent variable vary with i and t – see, respectively, (5), and (8) in the case of time-invariant covariates. *A fortiori*, the same can be said in the case of time-varying covariates. With regard to error correlation, correct specification of the conditional mean ensures that the autoregressive model is dynamically complete and, hence, its error term is serially uncorrelated. The same can obviously be ascertained with respect to the marginal model.

Under a marginal approach the expression of the skedastic function is not easy to deduce – see (8) – so the robust "sandwich" covariance estimator should be used for inference purposes, upon pooled NLS estimation of the marginal expectation parameters. On the other hand, under the autoregressive approach the expression of the conditional variance $V\left(y_{it}|y_{i,t-1},\boldsymbol{x}_{it}\right)$ is easy to obtain from (4) and (5); therefore, in this case one can also implement the asymptotically more efficient pooled weighted NLS estimator. In general, for time t (omitting i),

$$V\left(u_t|\boldsymbol{x}_t,y_{(t-1)}\right)=V\left(y_t|\boldsymbol{x}_t,y_{(t-1)}\right)=E\left(y_t^2|\boldsymbol{x}_t,y_{(t-1)}\right)-E^2\left(y_t|\boldsymbol{x}_t,y_{(t-1)}\right)=$$
$$\left(1-p_t\right)\left(p_t+\left(2p_t+p_{1t}\right)\left(1-p_{1t}\right)y_{t-1}+p_t\left(1-p_{1t}\right)^2y_{t-1}^2\right),$$

estimable by plugging in first-step pooled NLS estimates of the regression parameters.

3.2 Pooled Quasi-Maximum Likelihood

The QML method proceeds through maximization of a linear exponential family (LEF) likelihood. As is well known, the QML estimator is consistent and asymptotically normal regardless of the true likelihood, provided the means of the response variables are correctly specified (Gouriéroux, Monfort and Trognon, 1984). Therefore, one seemingly appealing route is to choose from several candidate univariate densities in the LEF family and estimate the parameters of the model by pooled QML.

Like NLS, for $n\to\infty$ QML is asymptotically valid under both a marginal and an autoregressive approach (see Appendix). As with pooled NLS, an autoregressive approach is clearly easier than a marginal approach when time-varying covariates are included in p and/or p_1.

The pooled QML estimator maximizes the quasi-log-likelihood $\sum_{i=1}^{n}LL_i$, with $LL_i\equiv\sum_{t=1}^{T_i}\log f\left(y_{it};\mu_{it}\right)$ and $f\left(\cdot;\cdot\right)$ denoting a particular LEF conditional density. Correct specification of the mean, μ_{it}, as a function of covariates, ensures that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{QML}-\boldsymbol{\theta}_0\right)\xrightarrow{d}\mathcal{N}\left(0,\boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}\right),$$

with $\hat{\boldsymbol{\theta}}_{QML}$ denoting the QML estimator of the parameters in p and p_1, $\mathcal{N}(\cdot,\cdot)$ the multivariate normal distribution, $\boldsymbol{\theta}_0$ the population parameter values and

$$\begin{aligned} \boldsymbol{A}_0 &\equiv E\left(-\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} LL_i\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \\ \boldsymbol{B}_0 &\equiv E\left(\nabla_{\boldsymbol{\theta}} LL_i \nabla_{\boldsymbol{\theta}'} LL_i\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \end{aligned} \tag{14}$$

Consistent estimators for A_0 and B_0 are obtained in the usual manner, replacing $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}_{QML}$ and population expectations by sample averages.(4)

In the present context natural choices for the quasi-likelihood are, among other possibilities, such discrete p.f.'s as the Poisson or negative binomial. In these cases, the density is given by, respectively,

$$\text{Poisson} \quad : \quad f^P(y_{it};\mu_{it}) = \exp(-\mu_{it})\,\mu_{it}^{y_{it}}/(y_{it}!)\,, \tag{15}$$

$$\text{Negative binomial} \quad : \quad f^{NB}(y_{it};\mu_{it}) = \frac{\Gamma(a+y_{it})}{\Gamma(a)\,\Gamma(y_{it}+1)}\left(\frac{\mu_{it}}{a}\right)^{y_{it}}\left(\frac{\mu_{it}}{a}+1\right)^{-(y_{it}+a)}, \tag{16}$$

where a denotes some given positive constant.(5) For instance, $a=1$ yields $f^{NB}(y_{it};\mu_{it}) = \mu_{it}^{y_{it}}(\mu_{it}+1)^{-(y_{it}+1)}$.

3.3 Maximum Likelihood

Under the assumption of independent individual time sequences ML estimation requires the joint probabilistic model for the Markov chain $\boldsymbol{y} \equiv (y_1,\ldots,y_T)$. This model is now made fully explicit in order to implement the ML approach.

In the present context $y_0 \equiv 0$ so $y_1 \equiv d_1$ and $\Pr(y_1) = p^{y_1}(1-p)^{1-y_1}$, with $y_1 \in \{0,1\}$. Assumption (1) yields the transition probabilities (2), now rewritten as (individual index omitted)

$$\Pr(y_t|y_{t-1}) = p^{\mathbf{1}(y_t=y_{t-1}+1)}\left((1-p)\binom{y_{t-1}}{y_t}p_1^{y_t}(1-p_1)^{y_{t-1}-y_t}\right)^{\mathbf{1}(y_t\le y_{t-1})},$$

[4] Inference should be conducted by using the corresponding covariance matrix estimator of $\hat{\boldsymbol{\theta}}_{QML}$ in the construction of standard errors and Wald statistics. A likelihood ratio-type statistic is not valid because of a very probably incorrect variance assumption, as well as neglected time dependence implied by the adopted likelihood.

[5] With a a positive constant, f^{NB} corresponds to a member of what is usually termed the NB2 model in the count data literature (where the conditional mean is usually specified as $\exp(\boldsymbol{x}_{it}'\boldsymbol{\beta})$ – see, e.g., Cameron and Trivedi, 2013).

where $\mathbf{1}(\cdot)$ denotes the usual indicator function. Then, with constant p and p_1 the joint conditional density of each individual sequence can be written as

$$\begin{aligned}
& f_y(\boldsymbol{y}|p,p_1) \\
= & \prod_{t=1}^{T}\left(p^{\mathbf{1}(y_t=y_{t-1}+1)}\left((1-p)\binom{y_{t-1}}{y_t}p_1^{y_t}(1-p_1)^{y_{t-1}-y_t}\right)^{\mathbf{1}(y_t\leq y_{t-1})}\right) \qquad (17)\\
= & \prod_{t=1}^{T}\left(\left(\frac{p}{1-p}\right)^{\mathbf{1}(y_t=y_{t-1}+1)}(1-p)\left(\binom{y_{t-1}}{y_t}\left(\frac{p_1}{1-p_1}\right)^{y_t}(1-p_1)^{y_{t-1}}\right)^{\mathbf{1}(y_t\leq y_{t-1})}\right)\\
\propto & \; p^{\sum_{t=1}^{T}\mathbf{1}(y_t=y_{t-1}+1)}(1-p)^{\sum_{t=1}^{T}\mathbf{1}(y_t\leq y_{t-1})}\times\\
& \left(\left(\frac{p_1}{1-p_1}\right)^{\sum_{t=2}^{T}\mathbf{1}(y_t\leq y_{t-1})y_t}(1-p_1)^{\sum_{t=2}^{T}\mathbf{1}(y_t\leq y_{t-1})y_{t-1}}\right)^{\mathbf{1}(T\geq 2)},
\end{aligned}$$

exhibiting, as expected, the usual split of hurdle models in two separate components. The first, involving p, refers to the binary process that splits individual sequences into 'success' (periods for which $y_t = y_{t-1}+1$), and 'failure' (periods for which $y_t \leq y_{t-1}$); the second, involving p_1, refers to the binomial part of the model, for periods with $y_t \leq y_{t-1}$.

ML estimates can be obtained on the basis of an individual contribution to the log-likelihood of the form

$$\begin{aligned}
LL_i = & \; const. + \sum_{t=1}^{T_i}\mathbf{1}(y_{it}=y_{i,t-1}+1)\log p + \sum_{t=1}^{T_i}\mathbf{1}(y_{it}\leq y_{i,t-1})\log(1-p) + \qquad (18)\\
& \mathbf{1}(T_i\geq 2)\left(\sum_{t=2}^{T_i}\mathbf{1}(y_{it}\leq y_{i,t-1})\left(y_{it}\log\frac{p_1}{1-p_1}+y_{i,t-1}\log(1-p_1)\right)\right).
\end{aligned}$$

It is readily seen that estimation of the first component of the hurdle (p estimation) uses all observations in the sample. Estimation of the second component (involving p_1) disregards data on the first period for every individual (consequently disregarding individuals with only one observation), as well as observations for periods with $d_{it} = 1$ so $y_{it} = y_{i,t-1}+1$.

In the case of a regression model, with covariates introduced in p and/or p_1, these should be appropriately specified and notation modified accordingly. With time-varying covariates – consider p_t and p_{1t} – (17) should be replaced by

$$\prod_{t=1}^{T}\left(\left(\frac{p_t}{1-p_t}\right)^{\mathbf{1}(y_t=y_{t-1}+1)}(1-p_t)\left(\binom{y_{t-1}}{y_t}\left(\frac{p_{1t}}{1-p_{1t}}\right)^{y_t}(1-p_{1t})^{y_{t-1}}\right)^{\mathbf{1}(y_t\leq y_{t-1})}\right) \qquad (19)$$

and the log-likelihood subsequently altered.

3.4 Unobserved Individual Heterogeneity

Allowing for unobserved individual heterogeneity within the present framework raises some specification and estimation issues that deserve caution. Frequently, in discrete choice models unobservables are added to the index functions within choice probabilities; for instance, with time-invariant unobserved effects, $e_{it} = e_i, \forall t$, and logit specifications one can have $p_{it} = \left(1 + \exp\left(-\boldsymbol{x}'_{it}\boldsymbol{\beta} - e_i\right)\right)^{-1}$ and similarly for p_{1it}.

Assume that $\boldsymbol{x}_i \equiv (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{T_i})$ are strictly exogenous conditionally on individual effects – formally, $E\left(y_{it}|\boldsymbol{x}_i, e_i\right) = E\left(y_{it}|\boldsymbol{x}_{it}, e_i\right)$. This condition agrees with the absence of lagged dependent variables in the definition of $\boldsymbol{x}_i$; also, the condition rules out feedback from y_t to future explanatory variables. Now, with e_i assumed independent of exogenous covariates a random effects approach involves estimation of $E\left(\boldsymbol{y}_i|\boldsymbol{x}_i\right) = E_e\left(E\left(\boldsymbol{y}_i|\boldsymbol{x}_i, e_i\right)\right)$, given some appropriate model for the inner expectation. This marginalization of heterogeneity likely involves integrals with no analytical solution, so estimation will entail the use of approximation techniques such as quadrature or Monte Carlo integration. Note, in addition, that this approach is only valid to estimate the parameters of the 'marginal' model (because e_i is not independent of lags of the dependent variables); therefore, with time-varying covariates in p and/or p_1 its application can become seriously burdensome.

Under the restriction that unobservables enter additively in p alone, a fixed effects strategy might also be used. The first component of the present hurdle specification is a binary model that uses data on $d_{it} \equiv \mathbf{1}\left(y_{it} = y_{i,t-1} + 1\right)$ alone – check (18). Thus, with p specified as logit, conditional ML, given a sufficient statistic for e_i, can be used to estimate its parameters – appropriate references in this regard are Chamberlain (1980) (no lagged responses in p), Chamberlain (1985) (dynamic pure time series logit model), or Honoré and Kyriazidou (2000) (other regressors in p, besides lagged responses). The functional separability of the hurdle likelihood then enables estimation of p_1 parameters through maximization of the second part of (18). The approach, however, is limited as it excludes unobservables from p_1 (not to mention computational complexity, even for not too large T – see, *e.g.*, Cameron and Trivedi, 2005, Ch. 23.4).[(6)]

[6] One alternative to the additive model of heterogeneity would be offered by a multiplicative model – as in $E\left(y_{it}|e_i\right) = e_i E\left(y_{it}\right)$, $t = 1, \ldots, T_i$. Possibly, this would allow the use of such fixed effects methods as Hausman, Hall and Griliches' (1984) conditional (multinomial-based) ML or Wooldridge's (1997) GMM estimators. However, in the present two-part model this assumption deserves caution as it rests on somewhat intricate and behaviourally odd assumptions regarding how heterogeneity affects

One viable random effects alternative strategy is to consider, in each period, p_t and/or p_{1t} random, with some appropriate conditional density, given covariates $\boldsymbol{x}_t$, and $\boldsymbol{y}_{(t-1)}$. One possibility for the latter is provided by the beta distribution, often used as a mixing density for probabilities. Here, p_t and/or p_{1t} can be specified as beta random variates with parameters that may depend on covariates; formally, for p_t (individual index omitted),

$$f\left(p_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right) = \frac{\Gamma\left(1/\alpha_t + 1/\left(\alpha_t\lambda_t\right)\right)}{\Gamma\left(1/\alpha_t\right)\Gamma\left(1/\left(\alpha_t\lambda_t\right)\right)} p_t^{1/\alpha_t - 1}\left(1-p_t\right)^{1/(\alpha_t\lambda_t)-1},$$

where α_t and λ_t denote positive parameters that may depend on $\boldsymbol{x}_t$ and $\boldsymbol{y}_{(t-1)}$. Analogously for p_1, with $f_1\left(p_{1t}|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$, involving positive parameters α_{1t} and λ_{1t} possibly depending on covariates as well. This general idea extends the approach of Santos Silva and Murteira (2009) for cross sectional single-part count data with bounded support, in which case it leads to the well known beta-binomial conditional p.f. (see also Heckman and Willis, 1977, for a well known seminal proposal of the beta-binomial p.f.).

The two-part framework naturally raises the possibility of statistical dependence of p_t and p_{1t}, as individual heterogeneity in each of p and p_1 is bound to be mutually dependent – a type of concern raised, *e.g.*, in Winkelmann (2004). While this concern is justified, allowing for dependence among p_t and p_{1t} (*e.g.* using copulas – see Joe, 1997) can render estimation more difficult (seemingly requiring approximation techniques), even under an autoregressive approach. Indeed, in this case, $E\left(y_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right) = E_{p,p_1}\left(p_t + r_t y_{t-1}|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$, the argument of which is nonlinear in p_t and p_{1t} – recall $r_t \equiv p_t + p_{1t}\left(1-p_t\right)$. Nonetheless, a not too strong assumption considerably facilitates estimation: if p_t and p_{1t} are assumed uncorrelated conditionally on $\boldsymbol{x}_t$ and $\boldsymbol{y}_{(t-1)}$, that is, $E\left(p_t p_{1t}|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right) = E\left(p_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right) E\left(p_{1t}|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$, then $E\left(y_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)$ can be easily obtained because $E_{p,p_1}\left(r_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right) = E\left(p_{1t}|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)\left(1 - E\left(p_t|\boldsymbol{x}_t, \boldsymbol{y}_{(t-1)}\right)\right)$. Note that this assumption does not rule out dependence among p_t and p_{1t}: nonlinear dependence conditional on $\boldsymbol{x}_t$ and $\boldsymbol{y}_{(t-1)}$ is fully allowed for, as well as marginal (to $\boldsymbol{x}_t$ and $\boldsymbol{y}_{(t-1)}$) dependence.

Under beta conditional densities for p_t and p_{1t}, then $E\left(p_t\right) = \lambda_t/\left(1+\lambda_t\right)$ and $E\left(p_{1t}\right) = \lambda_{1t}/\left(1+\lambda_{1t}\right)$; assuming p and p_1 are conditionally uncorrelated, one has (conditioning on

both p and p_1 in each period. In particular, an assumption of time-invariant individual heterogeneity affecting the conditional expectation of y_{it} rests on a time-varying unobservable term affecting p_1 (see the Appendix).

explanatory variables omitted)

$$E\left(y_t|\boldsymbol{x}_t,\boldsymbol{y}_{(t-1)}\right)=E_{p,p_1}\left(p_t+r_ty_{t-1}\right)=E\left(p_t\right)+E\left(p_{1t}\right)\left(1-E\left(p_t\right)\right)y_{t-1}=$$
$$\frac{\lambda_t}{1+\lambda_t}+\frac{\lambda_{1t}}{1+\lambda_{1t}}\left(1-\frac{\lambda_t}{1+\lambda_t}\right)y_{t-1}=\frac{\lambda_t}{1+\lambda_t}+\frac{\lambda_{1t}}{1+\lambda_{1t}}\frac{y_{t-1}}{1+\lambda_t}. \tag{20}$$

With a suitable parameterization of λ_t and λ_{1t} this expression enables estimation of the parameters of the autoregressive conditional mean of y_t, through pooled NLS or QML. In general, $\lambda=F\left(\boldsymbol{z}\right)/\left(1-F\left(\boldsymbol{z}\right)\right)$ ($\boldsymbol{z}$: covariates) yields $E\left(p_t\right)=F\left(\boldsymbol{z}\right)$, and similarly for p_{1t}. Hence, it seems appealing to specify λ in this way, by choosing some c.d.f. for F such as, *e.g.*, logit, probit or any other appropriate specification (other suggestions can be found, *e.g.*, in Ramalho, Ramalho and Murteira, 2011). In general, what this implies is that under correct specification of the conditional means of p_t and p_{1t}, pooled NLS and QML are valid estimators both without heterogeneity and under (beta distributed, conditionally uncorrelated) individual heterogeneity. Which means, in turn, that the estimation of the model through these methods – simpler than, *e.g.*, maximum simulated likelihood (as for the additive random effects approach described above) – may be more robust than at first envisaged. Further, in view of the known flexibility of the beta density, able to accomodate diverse heterogeneity patterns, this statement is likely to add up to the practical attractiveness of the general, beta-based, random effects approach.

Note, incidentally, that the marginal model can be obtained as before (without unobserved heterogeneity) – with the obvious caveat for analytical/computational intractability, if λ_t and λ_{1t} involve time-varying covariates. On the other hand, the likelihood function is no longer given by (17) or (19) but by the corresponding expectation with respect to the joint distribution of the T_i-vector of p_t and p_{1t}, $t=1,...,T_i$. Therefore, the estimator based on maximization of any of the former likelihoods is no longer valid under the suggested type of unobserved individual heterogeneity.

3.5 Specification Analysis

The specification of the proposed models depends crucially on the correct modelling of p and p_1, as functions of regressors. One feasible route to assess the specification of these conditional probabilities is provided by a RESET-type procedure (Ramsey, 1969; Pagan and Vella, 1989). In the present context, the test can be carried out by assessing the significance of powers of covariates' indices, as additional regressors within p and/or p_1.

Specifically, with logit p and p_1 the test can be implemented by assessing the significance of $(\boldsymbol{x}'_{it}\boldsymbol{\beta})^2$ and $(\boldsymbol{z}'_{it}\boldsymbol{\gamma})^2$ in

$$p_{it} \equiv \left(1+\exp\left(-\boldsymbol{x}'_{it}\boldsymbol{\beta}-(\boldsymbol{x}'_{it}\boldsymbol{\beta})^2\beta\right)\right)^{-1}, \quad p_{1it} \equiv \left(1+\exp\left(-\boldsymbol{z}'_{it}\boldsymbol{\gamma}-(\boldsymbol{z}'_{it}\boldsymbol{\gamma})^2\gamma\right)\right)^{-1}.$$

Obviously, additional powers can also be included and tested – see *e.g.* Ramalho and Ramalho (2012) for a comparative Monte Carlo study of several variants of the RESET test in binary choice models.

One other issue of particular interest in the present context regards the possibility of a single-part DGP, that is, the hypothesis that zeros and positive values of the responses are generated by the same conditional law. In this sense, one can reasonably assume that, under the null hypothesis, y_t given y_{t-1} follows a binomial p.f. with number of Bernoulli trials given by $y_{t-1}+1$. This hypothesis can be tested against the two-part specification as follows. Consider, firstly, a pure time series framework and suppose that under H_0 the probability of success is p_B, $0 \leq p_B \leq 1$, so $E\left(y_{it}|y_{i,t-1}\right) = \left(y_{i,t-1}+1\right)p_B$. Consequently, H_0 can be indirectly tested by assessing the significance of the intercept in the OLS regression

$$y_{it} = c + \left(y_{i,t-1}+1\right)p_B + error. \tag{21}$$

Failure to accept $H_0 : c = 0$ implies rejection of H_0, so a two-part model should reasonably be entertained.

In a regression framework, with p_B a specified function of observed regressors (*e.g.*, logit), the test can be implemented upon NLS estimation of the latter regression. Nonetheless, the simpler OLS procedure can also be used in the particular situation where p_B only involves strictly exogenous stationary covariates, so that $E\left(\boldsymbol{x}_t|y_{(t-1)}\right) = E\left(\boldsymbol{x}_t\right) = \boldsymbol{\mu}_{\boldsymbol{x}}$, $\forall t$, (which rules out feedback from previous values of the dependent variable to current explanatory variables). In this case, under H_0,

$$E\left(y_{it}|y_{i,t-1}\right) = E_{\boldsymbol{x}|y_{t-1}}\left(E\left(y_{it}|y_{i,t-1},\boldsymbol{x}_{it}\right)\right) = \left(y_{i,t-1}+1\right)E_{\boldsymbol{x}|y_{t-1}}\left(p_B\left(\boldsymbol{x}_{it}\right)|y_{i,t-1}\right) =$$

$$\left(y_{i,t-1}+1\right)E_{\boldsymbol{x}}\left(p_B\left(\boldsymbol{x}_{it}\right)\right) = \left(y_{i,t-1}+1\right)\times const.$$

Consequently, the test of H_0 can again be implemented upon simple OLS estimation of (21) – which is advantageous because it does not require previous specification of p_B as a function of covariates.[7] The next Section presents a succinct simulation exercise

[7] Note that if p_B involves lags of the dependent variable, *e.g.* $p_B = p_B\left(x_t, y_{t-1}\right)$, then $E\left(y_t|y_{t-1}\right) = \left(y_{t-1}+1\right)E_{\boldsymbol{x}|y_{t-1}}\left(p_B\left(\boldsymbol{x}_t, y_{t-1}\right)\right) = g\left(y_{t-1}\right)$, with $g\left(\cdot\right)$ some (generally nonconstant) function.

which assesses the finite sample performance of this procedure under single- and two-part DGP's.

The previous Section suggests a Hausman-type test to detect the presence of (conditional beta) unobserved heterogeneity. Under the null hypothesis of no heterogeneity both ML and pooled NLS/QML are $\sqrt{n}$-consistent estimators; however, in the presence of such heterogeneity, only the latter are consistent. Therefore, the null hypothesis to be tested is $\sqrt{n}$-consistency of both estimators. Under this hypothesis the ML estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{ML}$, is fully efficient (since it cannot be consistent otherwise, the likelihood not being a member of the LEF family); then, as is well known, the Hausman test has the form

$$n\left(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{ML}\right)'\left(\boldsymbol{V}_P-\boldsymbol{V}_M\right)^{-1}\left(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{ML}\right),$$

where $\tilde{\boldsymbol{\theta}}$ denotes either the pooled NLS or pooled QML estimator, $\boldsymbol{V}_P \equiv V\left(\sqrt{n}\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right)$ and $\boldsymbol{V}_M \equiv V\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}\right)\right)$. Under the null hypothesis $p\lim_{n\to\infty}\sqrt{n}\left(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{ML}\right)=\mathbf{0}$ this statistic is asymptotically distributed as a chi-squared random variate with number of degrees of freedom equal to the rank of $\boldsymbol{V}_P-\boldsymbol{V}_M$.(8)

4 Simulation Study

4.1 Performance of Alternative Estimators

Design

The performance of the pooled NLS, pooled QML and ML estimators of the conditional mean parameters are now compared on the basis of simulated data sets. These sets contain information on each individual time series of counts up to the sampling date, as well as on a set of covariates.

The value of the dependent variable for individual i at period t is denoted by y_{it}, where $1 \leq t \leq T_i$, and T_i represents the length of the i-th series up to the observation date (for instance, T_i can be measured in months). Throughout the exercise two samples were generated: the first with $n = 360$ independent individual time sequences and the second with $n = 5400$ sequences. In each sample $1 \leq T_i \leq 36$, with the same number of

[8] In practice the sample estimate for the covariance matrix to be used with the Hausman test may not be invertible, nor positive semi-definite, even if invertible. See Lee (1996, Ch. 5.9) for an estimator of the covariance matrix of $\sqrt{n}\left(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{ML}\right)$ that is positive semi-definite, by construction.

individuals for each different T_i value. That is, the smaller sample contains 10 (= 360/36) time series per each T_i, whereas the larger sample contains 150 (= 5400/36) time series per T_i.

Under the first type of DGP's (DGP1 and DGP2) no individual effects are considered, so individual sequences $\{y_{it}, t = 1, ..., T_i\}$ are drawn independently (independence across i) from the joint (conditional) p.f. defined in (17) with p and p_1 specified as logits. These include the following regressors and corresponding marginals: $x_1 \equiv 1$ (intercept); $x_2 \sim$ *Bernoulli* (.25); and $x_3 \sim \mathcal{N}(0, 1)$. These covariates are supposed time-invariant, so $x_{kit} \equiv x_{ki}$, $k = 2, 3$; in addition, a lagged dependent variable can be allowed for in p. Specifically, p and p_1 are specified as

$$p_{it} \equiv \left(1 + \exp\left(-\boldsymbol{x}_i'\boldsymbol{\beta} - \varsigma y_{i,t-1}\right)\right)^{-1}, \qquad p_{1i} \equiv \left(1 + \exp\left(-\boldsymbol{z}_i'\boldsymbol{\gamma}\right)\right)^{-1}, \tag{22}$$

with $\boldsymbol{x}_i \equiv (1, x_{i2}, x_{i3})'$, $\boldsymbol{z}_i \equiv (1, x_{i3})'$, $\boldsymbol{\beta} \equiv (.85, -.75, .25)'$, $\boldsymbol{\gamma} \equiv (1, -.5)'$ and $\varsigma = 0$ (DGP1) or $\varsigma = -.3$ (DGP2).

The last DGP, DGP3, involves unobserved individual heterogeneity. Under this process individual sequences $\{y_{it}, t = 1, ..., T_i\}$ are drawn independently according to the structure defined in (1), where, in each period, p_{it} now denotes a beta r.v. with parameters 1 and $1/\lambda_{it}$, where $\lambda_{it} = \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta} - .3y_{i,t-1}\right)$, and p_{1i} denotes a beta r.v. with parameters 1 and $1/\lambda_{1i}$, where $\lambda_{1i} = \exp\left(\boldsymbol{z}_i'\boldsymbol{\gamma}\right)$. This parameterization yields conditional means of p_{it} and p_{1i} as in (22). The covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ and the values of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are as in DGP2.[(9)] Under DGP3 the generated probabilities p_{it} and p_{1i} are conditionally uncorrelated, given $\boldsymbol{x}_i$ and $\boldsymbol{y}_{i(t-1)}$, so the conditional mean of y_{it} is given by (20), which, given the adopted parameterization of the conditional betas, yields logit functions for the expectations of p_{it} and p_{1i}.

The foregoing simulated data sets can be thought of as, *e.g.*, samples of loan repayment histories in fixed installments. Each individual represents a different contract aged T_i months at the sampling date (completed and incomplete contracts alike can be dealt with). The assumption of time-invariant regressors may reflect the frequent fact that clients' characteristics are recorded at the time of loan applications and remain subsequently

[9] In order to allow for different degrees of variability in p and p_1 the first parameter of the beta densities was initially allowed to vary with covariates, according to one of the functions $\exp(\delta x_{i3})$ and $\delta \in \{0, .1, .25, 1\}$. Given that the corresponding simulation results did not display noticeable differences, only those referring to $\delta = 0$ are included in the simulation results (Tables 3 and 6).

unchanged over the repayment period. A negative value for ς (p_{it} decreasing in $y_{i,t-1}$) can be naturally interpreted as a decrease in the probability *not* to pay, in response to an increase in the number of previously missed installments.

Under DGP1 and DGP2 both the autoregressive and marginal models are estimated – respectively, (4) and (6), with p and p_1 correctly specified as logits. All three estimators (NLS, QML and ML) are used for the autoregressive model whereas the marginal model is estimated by pooled NLS and QML. Thi is because p involves a time-varying covariate (lagged response) under DGP2, so the marginal expectation is clearly burdensome to employ (see (11)) – whereas the autoregressive approach involves no more than $E\left(y_t|y_{t-1}, \boldsymbol{x}\right)$. For both types of models the pooled QML estimator is implemented with Poisson, (15), and negative binomial ((16) with $a = 1$) likelihoods. Under DGP3 pooled NLS and pooled QML are used to estimate the autoregressive model, with correctly specified logit functions for the expectations of p_{it} and p_{1i}.

The performance of the various estimators is assessed on the basis of 2000 replications of the described samples, with regressors newly drawn at each replica. All computations, in the present and next Subsections, were performed using TSP 5.0 (Hall and Cummins, 2005).

Simulation Results

The results of the experiment are included in Table 1 through Table 6, referring to estimation results of correctly specified models under DGP1 through DGP3. For each estimator the tables display the average (Avg) and empirical root mean squared error (RMSE) over 2000 replicas, for $n = 360$ (tables $1-3$) and 5400 (tables $4-6$).

Tables 1 – 3 about here

Tables 4 – 6 about here

Overall, simulation results concerning valid estimators under the various DGP's used in the study can be considered to concord with broad theoretical expectations. In addition, several particular results appear noteworthy. Although asymptotically more efficient, the ML estimator does not appear to behave significantly better – actually, it often proves worse – than competing estimators, namely for the parameters in p and the smaller sample size. In this respect, the pooled QML estimator, namely with a negative binomial likelihood, seems to behave fully satisfactorily. Given that it is easier to implement than ML, pooled QML may be a perfectly sound choice to estimate the parameters of interest (either of the autoregressive or the marginal model).

The choice among an autoregressive or a marginal approach (also feasible when no time-varying covariates appear in p and/or p_1) is obviously indifferent on asymptotic grounds (as evinced by the larger sample results – check Table 4, for NLS and pooled QML estimates). Nonetheless, namely with a smaller sample, a comparative study of both types of approaches seems to be helpful for practical purposes. In this regard, the results in Table 1 seem to indicate that estimation of the autoregressive model yields better results than estimation of the marginal model (both on average and on efficiency grounds – as suggested by the root mean squared error results). A type of finding that clearly adds up to the user-friendlier nature of an autoregressive aproach, which, as mentioned, can be utilized in a wider range of situations than the marginal model.

4.2 Test of One-part Null Binomial Model

Design

This Section illustrates the performance of two variants of the test of the one-part binomial null specification that is proposed in Section 3.5. Both versions of the test are based on the test of significance of the intercept in regression (21). In the first version (T1) the test is carried out upon the OLS regression of y_{it} on $y_{i,t-1}+1$ and intercept; the second version (T2) is implemented upon NLS estimation of (21), with $p_{Bi} \equiv (1+\exp(\alpha_1+\alpha_2 z_i))^{-1}$ and z_i a random covariate defined below.

Two types of DGP are considered: in order to assess the empirical size of the tests, the data, for each i and t, are sequentially generated according to a binomial with 'parameters' $y_{i,t-1}+1$ and $p_{Bi} \equiv (1+\exp(-1+.5z_i))^{-1}$, where z_i denotes a time-invariant normal variate with null mean and variance $\sigma_z^2 \in \{.01, .25, 1\}$. In order to evaluate the tests' empirical power, the data are generated from the hurdle model with probabilities p_i, used for DGP1 in the previous subsection, and $p_{1i} = p_{Bi}$. In this experiment one sample is generated, with $n = 360$ independent individual time series, $1 \leq T_i \leq 36$ and the same number of individuals for each different T_i. As before, the exercise is based on 2000 replications of the described samples, with regressors newly drawn at each replica.

Simulation Results

The results of the experiment are included in Table 7 which displays rejection percentages for the proposed tests under the null (binomial model) and alternative hypotheses (hurdle model), respectively. Given the poor results for the T1 test under H_0, grossly over rejected by this procedure, the table only presents results for the T2 test under H_1. While the former test seems clearly unreliable, the empirical size and power results displyed by T2 suggest that this version of the test can lead to reliable conclusions about the convenience of choosing among a one-part vs. a two-part specification.

Table 7 about here

5 An Empirical Illustration

The present Section illustrates the application of the proposed methodologies to a data set on personal loans granted by a Portuguese financial institution. The data consist of a sample of 98 clients who were either repaying loans in February 2013, or had finished repaying their loans some time before this date. The earliest loan in the sample was contracted in October 2004 and the first loan to be fully repaid was completed in July 2010. The data set comprises both completed and active loans: 62 loans in the sample were still active in February 2013. For each contract the sample contains the time series of missed monthly installments at the end of each month, over the duration of the loan or from its beginning up to February 2013 (for loans not fully repaid by this date).

The available data set also contains information on some characteristics of the loans and of the clients – these variables are described in Table 8. The covariate H is a dummy variable indicating whether the credit is used to buy a house ($H = 1$) or not ($H = 0$). In the sample, 30 loans are for home purchase and involve a mortgage, while the remaining 68 credits are used for personal consumption. The covariate $EFRATE$ (effort rate) is included as a measure of the client's borrowing capacity and is defined as the ratio of the amount of the fixed installment to the client's family total monthly net income (according to the annual IRS official statement). For some (few) clients the value of $EFRATE$ is very high, reflecting, on the one hand, the over relaxed credit granting practice before the global financial crisis that surfaced in 2008 and, on the other hand, the consideration, in the decision to grant credit, of some clients' additional sources of income, not disclosed for tax purposes. The dummy variable $FCRIS$ (financial crisis) takes the value 0 before July 2009 and 1 as of this date, from which the ripples of the financial crisis can reasonably be judged to have become stronger in the Portuguese financial and economic environment. The proportion of zero responses in the sample is considerably larger than the total number of positive values: in a total of 5455 observations, there are 5143 zeros, with an overall sample average of .357 missed payments and a strongly positive skew (sample skewness coefficient equal to 12).

Table 8 about here

The null hypothesis of a single-part binomial model was tested as described in Section 4.2, by assessing the significance of the intercept in regression (21), with $p_B = p_1$ (defined below). This form of the test, based on NLS estimation of (21), was preferred to the OLS version, in view of the apparent unreliability of the latter, as suggested by the Monte Carlo results (in Section 4.2). The available data yielded an NLS intercept estimate of .018, highly significant (t statistic: 3.743). This result casts a strong doubt upon the convenience of a single-part specification, so a two-part model was estimated.

The main estimation results of the hurdle regression model are included in Table 9. These results refer to pooled QML (Poisson and negative binomial with $a = 1$ in (16)) and ML estimates, with p and p_1 specified as logits. No convergence was achieved with the NLS method, which prevented the presentation of the corresponding estimates. This setback notwithstanding, a sound alternative seems to be offered by pooled QML, no more difficult to use than NLS with common econometrics packages.

Table 9 about here

The covariates included in the model are indicated in Table 8. While the ML estimates of the parameters in p appear somewhat close to corresponding QML estimates, the same does not seem true with regard to p_1 parameters' estimates. Table 9 reports the results of a Hausman-type test, confronting ML and QML results. In both cases the outcome of the test suggests a clearcut rejection of the null hypothesis, which, as mentioned in Section 3.5, can be indication of unobserved individual heterogeneity in the data, under which presence ML is not valid. Accordingly, the subsequent analysis refers to QML estimates, deemed more reliable than ML estimation results. All interaction terms included as regressors in p and p_1 were statistically nonsignificant, so the estimated models do not include these terms. With regard to p_1, all variables initially considered, but $CAPITAL$, turned out to be nonsignificant so only the latter was included in this probability. Thus, p ($\equiv p_{it}$) was taken as individual- and time-specific (with the inclusion of the covariate $FCRIS$) and p_1 ($\equiv p_{1i}$) individual-specific. RESET-type tests were carried out by assessing the significance of squares of covariates indices in each of p and p_1; the associated robust t statistics and corresponding p-values suggest that the adopted specifications for both p and p_1 can be taken as appropriate approximations to the true conditional probabilities (Poisson, additional squared term in p: $t = .300$; in p_1: $t = .263$; negative binomial, additional squared term in p: $t = .348$; in p_1: $t = .060$).

Referring now to QML results, all the parameters' estimates are noted to be statistically significant at the 1% level. As expected, positive estimates were obtained for the coefficients of $EFRATE$, $FAMILY$ and $FCRIS$, all reasonably supposed to have a negative impact on the probability of loan repayment. The H dummy, in turn, is estimated to have a positive influence on this probability, a result in accordance with the fact that personal consumption loans ($H = 0$) usually involve less stringent types of collateral than mortgage loans ($H = 1$). The explanatory variables $CAPITAL$ and $DURATION$ are (quite naturally) highly correlated in the sample (correlation coefficient 84.6%) but both exert a statistically significant influence on p (though small and of opposite sign). Meanwhile, as alluded to before, the covariate $DURATION$ is individually insignificant in p_1, which is negatively affected by $CAPITAL$. At face value the probability of missing overdue installments seems negatively influenced by the amount borrowed; given that the sample contains a relatively small number of positive observations of the response variables, these results concerning p_1 estimation should be viewed with caution.

In order to illustrate how these results affect the expected value of missed payments, marginal effects of a unit change of $EFRATE$ (unit: percentage point) were computed under, respectively, $H = 0$ and 1, and $FCRIS = 0$ and 1. The corresponding estimates were obtained from the average of (12) for all individuals in the sample, considering $t = 56$ (the average of the age of contracts in the sample) and plugging in QML estimates. Table

10 displays these results, along with the estimated differences of the mean value of missed installments for $H = 0$ and 1 (first difference of (6), denoted $\Delta_H E(y_t|x)$) under both $FCRIS = 0$ and 1.

Table 10 about here

The contents of this table suggest a few remarks. Overall, estimates of marginal effects computed upon QML-Poisson are invariably larger, in absolute value, than those obtained from QML-negative binomial. Nevertheless, as expected, no sign contradictions are found between both methods and, in addition, both yield estimates of marginal effects that are roughly ten times higher, in absolute value, under $FCRIS = 1$ than under $FCRIS = 0$. In what concerns the marginal effect of one more percentage point of $EFRATE$ on $E(y_t|x)$, although apparently weak, it is strongest for $H = 0$ and $FCRIS = 1$ (under both QML methods). This is somehow expected, as, on the one hand, loans for which $H = 0$ usually involve weaker guarantees and, on the other hand, the general hardship caused by a major financial crisis ($FCRIS = 1$) is bound to affect the dependability of individuals' and households' repayment behaviour. In the opposite situation ($H = 1$, $FCRIS = 0$), the marginal effect of $EFRATE$ seems negligible: the presence of a strong collateral involved in mortgage loans may well serve as a strong deterrent of default in this case. Regarding the marginal effects of H on $E(y_t|x)$, estimates clearly suggest that, on average, the differential effect of a mortgage is stronger under a financial crisis than without it. This result may reflect the fact that, again, the presence of a strong collateral somehow stabilizes the repayment behaviour of borrowers, an effect that can be particularly felt under financial stress.

6 Final Remarks

This paper presents a hurdle model for count data with excess zeros and bounded support, suggesting its estimation through pooled NLS and QML, and ML. A Monte Carlo study suggests that the QML estimator competes fairly well with the ML estimator, even with moderate sample sizes. While being considerably easier to use, the pooled QML method is often more robust than ML, only requiring correct specification of the first conditional moments of the response variables. This feature is particularly useful when the data are affected by some forms of unobserved heterogeneity, which invalidate a full information maximum lilkelihood approach. In addition to the simulation study presented, a real data set on personal loans granted by a Portuguese financial institution illustrates the potential usefulness of the suggested approach, as a model of individual repayment behaviour. In this sense, the suggested models can prove a useful tool, for instance, in credit and behavioural scoring analysis.

The proposed approach can be generalized in several directions, extending its application to situations related to the present empirical study. For instance, through a redefinition of the support of each y_t the model can be used to address early repayment or redemption, frequent in credit granting applications. Or, by adapting the notion of thinning to continuous dependent variables, loan contracts with variable installments can also be addressed. Among others, these extensions provide a natural ground for future analytical and applied research.

Acknowledgements

José Murteira gratefully acknowledges financial support from Fundacao para a Ciencia e a Tecnologia, through strategic grant PEst-OE/EGE/UI0491/2011.

Appendix

This appendix presents algebraic derivations of expressions for relevant moments of the marginal, joint and conditional distributions involved in the sequence $\{y_t, t = 1, ..., T\}$. Also included is a brief statement of asymptotic properties of the QML estimator with $n \to \infty$ and bounded T_i.

Section 2.1 — $COV\left(y_t, y_{t-k}\right) = r^k V\left(y_{t-k}\right)$.

Proof. Through successive application of the law of iterated expectations to (4), one can write the first conditional moment of y_t, given y_{t-k}, as

$$E\left(y_t | y_{t-k}\right) = p\left(1 + r + ... + r^{k-1}\right) + r^k y_{t-k}. \tag{23}$$

From

$$COV\left(y_t, y_{t-k}\right) = E\left(y_t\left(y_{t-k} - E\left(y_{t-k}\right)\right)\right) = E\left(E\left(y_t | y_{t-k}\right)\left(y_{t-k} - E\left(y_{t-k}\right)\right)\right)$$

and from (23), it follows that

$$\begin{aligned} COV\left(y_t, y_{t-k}\right) &= E\left(\left(p\left(1 + r + ... + r^{k-1}\right) + r^k y_{t-k}\right)\left(y_{t-k} - E\left(y_{t-k}\right)\right)\right) \\ &= r^k E\left(y_{t-k}\left(y_{t-k} - E\left(y_{t-k}\right)\right)\right) = r^k V\left(y_{t-k}\right). \end{aligned}$$

■

Section 2.2 — General expression of $E\left(y_t | \boldsymbol{x}\right)$ with time-varying covariates – $\boldsymbol{x} \equiv \left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\right)$ does not include lags of dependent variables).

Proof. Suppose that both p and p_1 do not involve lags of the dependent variable. Then,

$$E\left(y_1 | \boldsymbol{x}_1\right) = p\left(\boldsymbol{x}_1\right),$$

$$E\left(y_2 | \boldsymbol{x}_{(2)}\right) = E\left(E\left(y_2 | y_1, \boldsymbol{x}_{(2)}\right)\right) = p\left(\boldsymbol{x}_2\right) + r\left(\boldsymbol{x}_2\right) E\left(y_1 | \boldsymbol{x}_1\right) = p\left(\boldsymbol{x}_2\right) + r\left(\boldsymbol{x}_2\right) p\left(\boldsymbol{x}_1\right),$$

$$E\left(y_3 | \boldsymbol{x}_{(3)}\right) = E\left(E\left(y_3 | y_2, \boldsymbol{x}_{(3)}\right)\right) = p\left(\boldsymbol{x}_3\right) + r\left(\boldsymbol{x}_3\right) E\left(y_2 | \boldsymbol{x}_{(2)}\right) =$$

$$p\left(\boldsymbol{x}_3\right) + r\left(\boldsymbol{x}_3\right)\left(p\left(\boldsymbol{x}_2\right) + r\left(\boldsymbol{x}_2\right) p\left(\boldsymbol{x}_1\right)\right) = p\left(\boldsymbol{x}_3\right) + p\left(\boldsymbol{x}_2\right) r\left(\boldsymbol{x}_3\right) + p\left(\boldsymbol{x}_1\right) r\left(\boldsymbol{x}_2\right) r\left(\boldsymbol{x}_3\right).$$

A mathematical induction argument ensures the general result

$$\begin{cases} p\left(\boldsymbol{x}_1\right) & t = 1, \\ p_t + \sum_{j=1}^{t-1} p_j \left(\prod_{k=j+1}^{t} r_k\right), & t \geq 2, \end{cases}$$

which reduces to (6) when all covariates (therefore p and p_1) are time-invariant. ■

Section ?? — Consistency and $\sqrt{n}$-asymptotic normality of the QML estimator.

Proof. The properties of consistency and $\sqrt{n}$-asymptotic normality of the QML estimator are well known to result from the fact that the population values of the conditional mean parameters maximize $E\left(LL_i\left(\theta\right)\right)$ – see Gouriéroux, et al. (1984). The following merely evinces the main necessary condition for consistency of QML under an 'autoregressive' approach (as defined in the main text).

Assume random sampling from the cross section. The individual contribution to the LEF log-likelihood can be written as (individual index, i, omitted so $LL \equiv LL_i$, $T \equiv T_i$ and so forth)

$$LL = const. + \sum_{t=1}^{T}\left(a\left(\mu_t\right) + c\left(\mu_t\right)y_t\right),$$

where $a\left(\cdot\right)$ and $c\left(\cdot\right)$ are functions such that $\mu_t = -c'\left(\mu_t\right)^{-1}a'\left(\mu_t\right)$, $\mu_t = p_t + r_t y_{t-1}$ and $y_0 \equiv 0$. Let $\boldsymbol{x}_t$ now denote all covariates except lagged dependent variables; at the population value of θ, under $E\left(y_t|y_{t-1},\boldsymbol{x}_t\right) = \mu_t, \forall t$,

$$E\left(\left.\frac{\partial LL}{\partial\theta}\right|\boldsymbol{x}\right) = E\left(\left.\sum_{t=1}^{T}c'\left(\mu_t\right)\left(y_t-\mu_t\right)\nabla_\theta\mu_t\right|\boldsymbol{x}\right) =$$
$$\sum_{t=1}^{T}E_{y_{t-1}|\boldsymbol{x}}\left(c'\left(\mu_t\right)\left(E\left(y_t|y_{t-1},\boldsymbol{x}_t\right)-\mu_t\right)\nabla_\theta\mu_t\right|\boldsymbol{x}\Big) = 0,$$

which enables the consistency of the QML estimator. ■

Section 3.4 – Multiplicative unobserved effects.

Proof. Let $\boldsymbol{e}_i \equiv \left(e_{i1},\ldots,e_{i,T_i}\right)$ and suppose that the mean of y_{it} is affected by a time-invariant effect multiplicatively, that is, $E\left(y_{it}|\boldsymbol{e}_i\right) = e_i E\left(y_{it}\right)$ (for simplicity, time-invariant observable covariates are assumed so they are omitted). Hence, for $t = 1$,

$$E\left(y_{i1}|e_i\right) = e_i E\left(y_{i1}\right) = e_i \Pr\left(d_{i1} = 1\right) = e_i p.$$

For $t \geq 2$, conditionally on $d_{it} = 0$ and $y_{i,t-1}$, y_{it} follows a binomial p.f. with number of Bernoulli trials given by $y_{i,t-1}$, so an unobserved effect will intervene in the probability of 'success', p_1; denote this as $p_1\left(e_{1it}\right)$. Recalling (1), one can check that, for $t = 2$, the term e_{1i2} must satisfy the equation

$$E\left(y_{i2}|e_i\right) = e_i E\left(y_{i2}\right) \Leftrightarrow$$

$$e_i p\left(1 + e_i p + p_1\left(e_{1i2}\right)\left(1 - e_i p\right)\right) = e_i p\left(1 + p + p_1\left(1 - p\right)\right)$$

$$\Leftrightarrow$$

$$e_i p + p_1\left(e_{1i2}\right)\left(1 - e_i p\right) = p + p_1\left(1 - p\right).$$

This equation is formally different from those that are obtained for $t > 2$, so, for each t, the roots e_{1it} of the corresponding implicit equations vary with t. In words, an assumption of time-invariant individual heterogeneity affecting the conditional expectation of y_{it} rests on a time-varying unobservable term affecting p_1. ■

References

- Al-Osh, M.A., Alzaid, A.A. (1987). "First-order integer valued autoregressive INAR(1) process". *Journal of Time Series Analysis* 8: 261-275.

- Brännäs, K. (1994). "Estimation and testing in integer valued AR(1) models". *Umeå Economic Studies* 355. University of Umeå.

- Brännäs, K. (1995). "Explanatory variables in the AR(1) model". *Umeå Economic Studies* 381. University of Umeå.

- Cameron, A.C., Trivedi, P.K. (2005). *Microeconometrics Methods and Applications.* Cambridge: Cambridge University Press.

- Cameron, A.C., Trivedi, P.K. (2013). *Regression Analysis of Count Data, 2nd Ed.*. Cambridge: Cambridge University Press.

- Chamberlain, G. (1980). "Analysis of covariance with qualitative data". *Review of Economic Studies* 47: 225-238.

- Chamberlain, G. (1985). "Heterogeneity, omitted variable bias, and duration dependence". In: Heckman, J.J., Singer, B., Eds.. *Longitudinal Analysis of Labor Market Data.* Cambridge: Cambridge University Press, pp. 3-38.

- Cragg, J.G. (1971). "Some statistical models for limited dependent variables with application to the demand for durable goods". *Econometrica* 39: 829-844.

- Gourieroux, C., Monfort, A., Trognon, A. (1984). "Pseudo maximum likelihood methods: theory". *Econometrica* 52: 681-700.

- Hall, B.H., Cummins, C. (2005). *TSP 5.0 User's Guide.* Palo Alto (CA): TSP International.

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

- Heckman, J.J., Willis, R.J. (1977). "A beta-logistic model for the analysis of sequential labor force participation by married women". *Journal of Political Economy* 85: 27-58.

- Honoré, B., Kyriazidou, E. (2000). "Panel data discrete choice models with lagged dependent variables". *Econometrica* 68(4): 839–874.

- Jacobs, P. A., Lewis, P.A.W. (1978). "Discrete time series generated by mixtures II: asymptotic properties". *Journal of The Royal Statistical Society B* 40: 222-228.

- Jin-Guan, D., Yuan, L. (1991). "The integer-valued autoregressive (INAR(p)) model". *Journal of Time Series Analysis* 12: 129-142.

- Joe, H. (1997). *Multivariate Models and Dependence Concepts.* London: Chapman & Hall.

- Jung, R.C., Kukuk, M., Liesenfeld, R. (2006). "Time series of count data: modelling, estimation and diagnostics". *Computational Statistics and Data Analysis* 51: 2350–2364.

- Lee, M.-J. (1996). *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models.* New York: Springer.

- MacDonald, I.L., Zucchini, W. (1997). *Hidden Markov and OtherModels for Discrete-Valued Time Series.* London: Chapman & Hall.

- McKenzie, E. (1985). "Some simple models for discrete variate time series". *Water Resources Bulletin* 21: 645-650.

- McKenzie, E. (1988). "Some ARMA models for dependent sequences of Poisson counts". *Advances in Applied Probability* 22: 822-835.

- McKenzie, E. (2003). "Discrete variate time series, stochastic processes: modelling and simulation". In: Shanbhag, D.N., Rao, C.R., Eds., *Handbook of Statistics vol. 21.* Amsterdam: North-Holland, pp. 573-606.

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

- Pagan, A., Vella, F. (1989). "Diagnostic tests for models based on individual data: a survey". *Journal of Applied Econometrics* 4: 29-59.

- Ramalho, E.A., Ramalho, J.J.S., Murteira, J.M.R. (2011). "Alternative estimating and testing empirical strategies for fractional regression models". *Journal of Economic Surveys* 25: 19-68.

- Ramalho, E.A., Ramalho, J.J.S. (2012). "Alternative versions of the RESET test for binary response index models: a comparative study". *Oxford Bulletin of Economics and Statistics* 74: 107-130.

- Ramsey, J.B. (1969). "Tests for specification errors in classical linear least squares regression analysis". *Journal of the Royal Statistical Society B* 31: 350–371.

- Ronning, G., Jung, R.C. (1992). "Estimation of a first order autoregressive process with Poisson marginals for count data". In: Fahrmeir, L., Francis, B., Gilchrist, R., Tutz, G., Eds., *Advances in GLIM and Statistical Modelling.* New York: Springer.

- Santos Silva, J.M.C., Murteira, J.M.R. (2009). "Estimation of default probabilities with incomplete contracts data". *Journal of Empirical Finance* 16: 457–465.

- Steutel, F.W., VanHarn, K. (1979). "Discrete analogues of self-decomposability and stability". *Annals of Probability* 7: 893-899.

- Sun, J., Zhao, X. (2013). *Statistical Analysis of Panel Count Data.* New York: Springer.

- Thomas, L.C., Edelman, D.B., Crook, J.N. (2002). *Credit Scoring and its Applications.* Philadelphia: SIAM.

- Weiss, C. (2008). "Thinning operations for modeling time series of counts – a survey". *Advances in Statistical Analysis* 92: 319-341.

- Windmeijer, F. (2006). "GMM for Panel Count Data Models". *CeMMAP working papers CWP21/06.* Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

- Winkelmann, R. (2004). "Health care reform and the number of doctor visits – an econometric analysis". *Journal of Applied Econometrics* 19: 455-472.

- Wooldridge, J. (1997). "Multiplicative panel data models without the strict exogeneity assumption". *Econometric Theory* 13: 667-678.

Table 1 - Performance of Estimators, DGP1, $n = 360$

Autoregressive Model

	NLS		QML-Poisson		QML-Neg. Bin.		ML	
	Avg	RMSE	Avg	RMSE	Avg	RMSE	Avg	RMSE
β_1	.870	.108	.858	.070	.856	.072	.852	.032
β_2	-.756	.090	-.754	.069	-.754	.081	-.753	.059
β_3	.257	.121	.252	.064	.249	.069	.250	.027
γ_1	.957	.206	.985	.127	.992	.141	1.000	.024
γ_2	-.543	.244	-.514	.136	-.507	.155	-.501	.026

Marginal Model

	NLS		QML-Poisson		QML-Neg. Bin.			
	Avg	RMSE	Avg	RMSE	Avg	RMSE		
β_1	.874	.207	.863	.122	.861	.105	–	–
β_2	-.757	.081	-.755	.077	-.756	.075	–	–
β_3	.233	.174	.255	.118	.255	.097	–	–
γ_1	.581	5.078	.968	.234	.945	1.366	–	–
γ_2	-.487	.530	-.535	.255	-.532	.241	–	–

Table 2 - Performance of Estimators, DGP2, $n = 360$

Autoregressive Model

	NLS		QML-Poisson		QML-Neg. Bin.		ML	
	Avg	RMSE	Avg	RMSE	Avg	RMSE	Avg	RMSE
β_1	.886	.083	.866	.063	.861	.064	.819	.062
β_2	-.754	.105	-.751	.087	-.752	.093	-.736	.062
β_3	.259	.068	.255	.050	.251	.054	.249	.026
ς	-.307	.262	-.301	.246	-.303	.243	-.284	.201
γ_1	.950	.325	.957	.307	.973	.326	1.000	.026
γ_2	-.536	.119	-.528	.097	-.521	.104	-.499	.026

Table 3 - Performance of Estimators, DGP3, $n = 360$

Autoregressive Model

	NLS		QML-Poisson		QML-Neg. Bin.	
	Avg	RMSE	Avg	RMSE	Avg	RMSE
β_1	.634	.237	.599	.260	.589	.269
β_2	-.519	.251	-.503	.259	-.504	.260
β_3	.180	.098	.173	.092	.171	.095
ς	-.231	.309	-.207	.302	-.211	.309
γ_1	.631	.580	.581	.618	.613	.643
γ_2	-.020	.497	-.016	.495	-.010	.502

Table 4 - Performance of Estimators, DGP1, $n = 5400$

Autoregressive Model

	NLS		QML-Poisson		QML-Neg. Bin.		ML	
	Avg	RMSE	Avg	RMSE	Avg	RMSE	Avg	RMSE
β_1	.851	.027	.850	.017	.850	.018	.850	.008
β_2	-.749	.023	-.749	.017	-.749	.020	-.749	.015
β_3	.252	.030	.250	.016	.250	.018	.250	.007
γ_1	.997	.050	.999	.032	1.000	.035	1.000	.006
γ_2	-.506	.055	-.501	.033	-0.501	.039	-.500	.007

Marginal Model

	NLS		QML-Poisson		QML-Neg. Bin.	
	Avg	RMSE	Avg	RMSE	Avg	RMSE
β_1	.851	.042	.851	.031	.850	.025
β_2	-.749	.020	-.749	.019	-.749	.018
β_3	.253	.041	.252	.030	.252	.024
γ_1	.996	.078	.997	.058	.998	.048
γ_2	-.508	.080	-.506	.060	-.504	.049

Table 5 - Performance of Estimators, DGP2, $n = 5400$

Autoregressive Model

	NLS		QML-Poisson		QML-Neg. Bin.		ML	
	Avg	RMSE	Avg	RMSE	Avg	RMSE	Avg	RMSE
β_1	.852	.022	.851	.016	.851	.017	.817	.036
β_2	-.749	.032	-.750	.024	-.750	.025	-.736	.021
β_3	.251	.017	.251	.013	.251	.014	.250	.007
ς	-.299	.206	-.300	.204	-.300	.204	-.283	.200
γ_1	.992	.100	.996	.085	.997	.088	1.000	.006
γ_2	-.504	.031	-.503	.025	-.502	.027	-.500	.006

Table 6 - Performance of Estimators, DGP3, $n = 5400$

Autoregressive Model

	NLS		QML-Poisson		QML-Neg. Bin.	
	Avg	RMSE	Avg	RMSE	Avg	RMSE
β_1	.793	.075	.746	.083	.735	.084
β_2	-.791	.073	-.778	.077	-.762	.081
β_3	.223	.039	..209	.034	.211	.037
ς	-.274	.123	-.270	.145	-.258	.148
γ_1	.830	.201	.817	.228	.815	.231
γ_2	-.357	.151	-.337	.163	-.336	.161

Table 7

Tests of One-part Binomial Model

Rejection Percentages, nominal level 5%

$H_0 : y_{it}$ Binomial			
σ_z^2	.01	.25	1
T1	46.43	47.15	51.28
T2	3.32	3.87	4.07
$H_1 : y_{it}$ Hurdle Model			
σ_z^2	.01	.25	1
T2	93.27	95.34	95.28

Table 8 - Description of Variables Used in the Empirical Illustration

Definition	mean (sum)	st.dev.
$EFRATE$: (installment/monthly net income) × 100%	22.284%	18.880
$FAMILY$: Number of family members	2.867	1.375
$CAPITAL$: Total borrowed amount (1000 euros)	$34,285$	$44,653$
$DURATION$: Contract duration of return period (months)	157.153	161.385
$H = 1$, if loan is used to buy a house	(30)	
$FCRIS = 1$ as of July 2009		

Table 9 - Estimation Results and Hausman Tests

Estimator	ML			QML Poisson			QML Neg. Binom.		
	est.	s.e.	*pv*	est.	s.e.	*pv*	est.	s.e.	*pv*
p									
Intercept	−7.105	.425	.000	−7.539	.637	.000	−7.570	.641	.000
EFRATE	2.339	.120	.000	2.144	.421	.000	2.184	.430	.000
FAMILY	.315	.031	.000	.405	.084	.000	.412	.084	.000
CAPITAL	−.00004	.000006	.000	−.00003	.000007	.000	−.00003	.000007	.000
DURATION	.015	.002	.000	.010	.003	.003	.010	.003	.002
H	−4.051	.554	.000	−2.883	.923	.002	−2.901	.929	.002
FCRIS	2.218	.430	.000	2.447	.432	.000	2.470	.433	.000
p_1									
Intercept	.501	.106	.000	4.688	.649	.000	2.596	.313	.000
CAPITAL	−.00002	.000008	.003	−.00007	.00001	.000	−.00004	.00001	.000
Hausman Test				Poisson *vs.* ML			Neg. Bin. *vs.* ML		
				Test Statistic		*pv*	Test Statistic		*pv*
				60.544		≈ 0	64.919		≈ 0

Table 10

Average Marginal Effects with respect to $EFRATE$ and H

$\nabla_{EFRATE}E\left(y_t \mid x\right)$, $t = 56$			
$H = 0, FCRIS = 0$		$H = 0, FCRIS = 1$	
Poisson	Neg. Bin.	Poisson	Neg. Bin.
.0018	.0002	.0179	.0033
$H = 1, FCRIS = 0$		$H = 1, FCRIS = 1$	
Poisson	Neg. Bin.	Poisson	Neg. Bin.
.0001	.00001	.0012	.0001
$\Delta_H E\left(y_t \mid x\right)$, $t = 56$			
$FCRIS = 0$		$FCRIS = 1$	
Poisson	Neg. Bin.	Poisson	Neg. Bin.
-.1479	-.0482	-1.4488	-.5557