Modelling Panel Count Data With Excess Zeros: A Hurdle Approach*

José M. R. Murteira

Faculdade de Economia, Universidade de Coimbra, and CEMAPRE

March, 2007

Abstract

This paper presents a hurdle-type regression model for panel count data. Several specific features of the data generating process require a modelling approach that differs from some commonly used count data models. The suggested model is intended for count data with excess zeros, relative to simple Poisson generated data, and bounded support. This kind of data may occur, *e.g.*, in the context of consumer credit or behavioural scoring, with reference to the repayment of loans in a pre-determined number of fixed installments. At each repayment date a debtor faces a known number of missed installments which depends on his previous repayment decisions and is bounded by the age of the contract. In any case, as most clients 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, e.g., the Poisson distribution.

Throughout the paper a random sample of observations on time series of counts and some set of covariates is supposed to be available to the researcher. The dependence structure within each time series is addressed through the use of the "binomial thinning" operator (Steutel and Van-Harn, 1979). As detailed in the text, this operator provides a flexible way

*I am grateful to João Santos Silva for helpful comments. Support from Fundação para a Ciência e Tecnologia, program FEDER/POCI 2010, is also gratefully acknowledged. The usual disclaimer applies. Address: José M. R. Murteira, Faculdade de Economia, Universidade de Coimbra, Av. Dias da Silva, 165, 3004-512 Coimbra, Portugal. e-mail: jmurt@fe.uc.pt.

to model the dependence between consecutive counts, within a Markov chain-type modelling approach.

Estimation of the model through conditional least squares and maximum likelihood is illustrated on the basis of different simulated data sets of independent individual time sequences of counts. The performance of both methods appears somewhat similar, with slight expectable efficiency advantages of maximum likelihood in cases of correct model specification.

The proposed specification can be used, for instance, as a model of repayment behaviour, to be employed in a context of credit or behavioural scoring. In this sense it may provide better estimates of default probability than, *e.g.*, simple cross-section models of the total number of missed installments. Meanwhile, with some adjustments, the approach is rich enough to encompass such situations as early repayment or redemption of loans, as well as loan contracts with variable installments.

JEL classification: C23, C25

Key Words: panel count data; hurdle model; conditional least squares; maximum likelihood.

1 Introduction

This paper introduces a hurdle-type regression model for panel count data. Several specific features of the data generating process require a modelling approach that differs from some commonly used count data models. The proposed specification is intended for count data with excess zeros, relative to simple Poisson generated data, and bounded support. Such data may occur, for instance, in the context of consumer credit or behavioural scoring, with reference to the repayment of loans in a pre-determined number of fixed installments: at each repayment date a debtor faces a known number of missed installments, which depends on his previous repayment decisions and is bounded by the age of the contract. In any case, as most clients 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, *e.g.*, the Poisson distribution.

Throughout the text a random sample of observations on time series of counts and some set of covariates is supposed to be available to the researcher. For each statistical unit at each observation period the possibility of excess zeros, relative to the simple Poisson, is dealt with through a hurdle-type specification (Cragg, 1971, Mullahy, 1986). In addition, the inherently bounded support of the dependent count variable also excludes the Poisson or negative binomial, among others, from the set of candidate models to be selected. In the present context the binomial distribution or a binomial-based mixture seem clearly more appropriate than functions with unbounded support.

The specification proposed here is designed to model time series of counts, one for each statistical unit. The literature proposes many possible ways to model time series of discrete variables. *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 p.f.'s 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). Brännäs (1995) was the first to extend these models to the regression case. Several other approaches are discussed in the literature on discrete variables time series analysis: surveys can be found in Cameron and Trivedi (1998), ch. 7, and McKenzie (2003).

The dependence structure within each time series is addressed in the present case through the use of the "binomial thinning" operator (Steutel and VanHarn, 1979). This operator provides a flexible way to model the dependence between counts, with the observation for one period defining the support bound of the count variable for the following period. As detailed below, this feature seems quite adapt to the type of data the proposed model is intended for.

The paper is organized as follows. After specification of the general model in section 2, section 3 presents its estimation through conditional least squares (CLS) and maximum likelihood (ML). The performance of these methods is assessed in section 4 on the basis of different simulated data sets. Algebraic derivations and proofs are included in the final appendix.

2 Model Specification

Consider a sample of observations on n time series of counts, y_{it}, where $t = 1, ..., T_i$, and $i = 1, ..., n$ denote, respectively, time and individual observational indices. In what follows a model for each individual time sequence of counts is presented, enabling the specification of a panel count data model. Estimation and inference issues are discussed in the next section.

The proposed specification starts with the following structural assumption about the triplet (y_{it},$y_{i,t-1}$,d_{it}) (individual index, i, omitted):

$$y_t = d_t \left(y_{t-1} + 1\right) + \left(1 - d_t\right) \left(p_1 \circ y_{t-1}\right), \tag{1}$$
$$y_t \in \{0, 1, ..., t\}, \quad t = 1, ..., T,$$

where $y_0 \left(\equiv y_{i0}\right) \equiv 0$, $d_t \left(\equiv d_{it}\right) \in \{0, 1\}$ denotes a Bernoulli random variable with $\Pr\left(d_t = 1\right) = p$, 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 y represents a positive integer, then $p_1 \circ y \equiv \sum_{j=1}^{y} b_j \left(p_1\right)$, where $\{b_j \left(p_1\right), j = 1, ..., y\}$

denotes a set of i.i.d. Bernoulli random variables independent of y, for which $\Pr(b_j(p_1) = 1) = p_1$. That is, conditional on y, $p_1 \circ y$ is a binomial random variable, the number of successes in y independent trials in each of which the probability of success is p_1.

Thus, conditional on y_{t-1}, the support of y_t 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 y_t conditional on y_{t-1} follows a binomial p.f. with parameters y_{t-1} and p_1. The probabilistic model for $y_{it}|y_{i,t-1}$ 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.

Unless $p = 0$ (corresponding to the trivial case $y_t \equiv 0, \forall t$), the sequence $\{y_t\}$ is not stationary. This can be seen by considering the sequence of unconditional first moments of y_t,

$$E(y_t) = p\left(1 + r + r^2 + \cdots + r^{t-1}\right),$$

where $r \equiv p + p_1(1 - p)$.[1] The covariance function is given by

$$COV(y_t, y_{t-k}) = r^k VAR(y_{t-k}),$$

a similar result to the expression for the covariance function in linear AR models for continuous data. It follows that the autocorrelation function,

$$r^k \sqrt{VAR(y_{t-k}) / VAR(y_t)},$$

depends, not only on lag (k), but also on t (as $VAR(y_j)$ involves j).[2]

For each individual the proposed model can be cast within the general framework of a first-order Markov chain. For each individual time sequence of T_i

[1] For the trivial case $p_i = 1$ ($\Leftrightarrow r_i = 1$), one obtains

$$E(y_{it}|y_{i,t-1}) = y_{i,t-1} + 1 = t,\ VAR(y_{it}|y_{i,t-1}) = 0.$$

[2] The formula for $E\left(y_t^2\right)$ can be obtained as the solution to the difference equation

$$E\left(y_{t+1}^2\right) = \left(p + (r-p)^2/(1-p)\right) E\left(y_t^2\right)$$
$$+ (1-r)(r-p)\, p\left(1 + r + \cdots + r^{t-1}\right)/(1-p) + p,$$
$$t = 2, ..., l,$$

with initial condition $E\left(y_1^2\right) = p$. Obviously, the expression for $VAR(y_t)$ involves t.

count variables, $\{y_{it}, t = 1, ..., T_i\}$, one can formally write the following recurrence formula for $\Pr(y_{it})$ (individual index and possible conditioning covariates omitted):

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

with transition probabilities, $\Pr(y_t = y|y_{t-1} = j-1)$, obtained from (1) and $\Pr(y_0 = 0) = 1$.

For each t, the last recurrence formula can be written in matrix form as

$$\pi_t = M\pi_{t-1} = M^t\pi_0,$$

with

$$\underset{(T+1)\times 1}{\pi_0} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \underset{(T+1)\times 1}{\pi_t} = \begin{bmatrix} \Pr(y_t = 0) \\ \Pr(y_t = 1) \\ \vdots \\ \Pr(y_t = t) \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad t = 1, ..., T,$$

and transition matrix

$$\underset{(T+1)\times(T+1)}{M} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & \cdots & m_{1T} & 0 \\ m_{21} & m_{22} & m_{23} & \cdots & m_{2T} & 0 \\ 0 & m_{32} & m_{33} & \cdots & m_{3T} & 0 \\ 0 & 0 & m_{43} & \cdots & m_{4T} & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & m_{T+1,T} & 0 \end{bmatrix},$$

where $m_{jk} \equiv \Pr(y_t = j-1|y_{t-1} = k-1)$, $1 \leq k \leq T$, $1 \leq j \leq k+1$.

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

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

where $\{d_t\}$ are i.i.d. binary and $\{z_t\}$ are i.i.d. with given distribution. If y_0 is also sampled from this distribution, then this model defines a stationary process with that same marginal distribution. In model (1) not only stationarity is absent (because of the term $y_{t-1}+1$) but also the variable replacing z_t ($p_1 \circ y_{t-1}$) is not i.i.d.. In addition, the assumption of independence between d_t and $p_1 \circ y_{t-1}$ is to be relaxed below.

If one thinks of the individuals in the sample 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 $y_{it}|y_{i,t-1}$ can be seen as a model of repayment behaviour. d_t is defined to be zero if the client decides to pay at period t (that is, "success" refers to a nonpayment choice, with probability denoted by p). If 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 he has 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. First, 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.

The above example illustrates the possible meaning of several limiting (even if unlikely) cases concerning different values of p and p_1. First, $p = 0$ (d_t degenerate at zero, $\forall t$), means that the borrower always decides to pay, one installment at least. From $y_0 \equiv 0$ it follows that $y_1 \equiv d_1 = 0$; as $y_t = p_1 \circ y_{t-1}$ and the individual always pays, $y_{t-1} = 0 = y_t$, $\forall t$, 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.

Now consider the case $p_1 = 0$. Then, once the client decides to pay ($d_t = 0$) he pays all the missed payments at that time (the b_j are degenerate at zero, $j = 1, ..., t-1$), in addition to that period installment. That is, at each repayment date a client 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 (b_j degenerate at one, $j = 1, ..., t-1$). 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) clients who regularly meet their obligations, p can be positive due to the possible occurrence of random unexpected (and undesirable) events, affecting their financial capacity and hindering their plans to pay one installment after the other during the life of the contract. In any case, a well known advantage of the hurdle specification is that it enables 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.[3]

It may be more appropriate to consider p and p_1 individual-specific, rather than constant across different individuals (p_i and p_{1i} in (1), instead of p and p_1). Also, a pure time series model may not lead to very useful conclusions.[4]. In order to account for this individual heterogeneity one obvious possibility is to introduce regressors in the model (*e.g.* in p and p_1). Nevertheless, even within a regression approach unobserved individual and/or time effects may still be at play here, requiring a more sophisticated approach concerning model estimation. If this is the case, the obvious decision regards the choice between a fixed effects approach or a random effects approach, marginal to unobserved individual or time effects. As usual, the appropriate choice depends on the objective of the analysis. For instance, a random effects approach seems more adapt for credit scoring, usually involving a large number of individual loan contracts.

The assumption of independence about the joint probabilistic behavior of $(d_t, b_1, ..., b_{t-1}, y_{t-1})$ may 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 of this dilemma 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 (possibly

[3] An extended survey on credit and behavioural scoring techniques and applications can be found, for instance, in Thomas, Edelman and Crook (2002).

[4] For instance, credit or behavioural scoring purposes may require the classification of debtors on the basis of contracts and customers characteristics.

including lagged values of the dependent variable) and/or unobservable effects in the model, through p and p_1. For instance, p and p_1 can be specified as logits: with time-invariant regressor vectors x_i and z_i, p_i and p_{1i} can be expressed as

$$\begin{aligned} p_i &\equiv \left(1+\exp\left(-x_i'\beta\right)\right)^{-1}, \\ p_{1i} &\equiv \left(1+\exp\left(-z_i'\gamma\right)\right)^{-1}, \end{aligned} \tag{2}$$

where β and γ denote regressor vectors. Obviously, time-varying regressors can also be introduced – then one would naturally use p_{it} and p_{1it}, instead. Unobserved heterogeneity can also be accounted for but, as expected, this concern can actually lead to increased difficulty in what regards model estimation.

3 Estimation and Inference

The present section discusses estimation and testing of the panel count data model introduced above. Two estimation methods are considered in what follows: CLS and ML. Use of both methods in the context of the stationary Poisson INAR is studied by Al-Osh and Alzaid (1987), Jin-Guan and Yuan (1991) and Ronning and Jung (1992). 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 CLS and GMM. A very recent account of estimating methods for panel count data, namely GMM, can be found in Windmeijer (2006).

3.1 Conditional Least Squares

Under the assumptions of the model and in view of the definition of the binomial thinning operator, the conditional mean of y_{it} can be written as

$$\begin{aligned} E\left(y_{it}|y_{i,t-1}\right) &= E\left(d_{it}\left(y_{i,t-1}+1\right)+\left(1-d_{it}\right)\left(p_{1i}\circ y_{i,t-1}\right)\right) \\ &= p_i + r_i y_{i,t-1}, \quad t=1,...,T_i, \end{aligned} \tag{3}$$

where $y_{i0} \equiv 0$ and $r_i \equiv p_i + p_{1i}\left(1-p_i\right)$. Consider, for now, a pure time series framework, with p and p_1 assumed constant and a panel of n independent individual time series, $\{y_{i1},...,y_{iT_i}\}$, $i=1,...,n$, (y_{it} independent of y_{ju}, $\forall i \neq$

$j, \forall t, u$). Then, the CLS estimator of the parameters p and p_1 minimizes the criterion function

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(y_{it}-p-ry_{i,t-1}\right)^2.$$

This objective function is the one associated with nonlinear least squares (NLS) estimation of the pooled regression model

$$\begin{aligned} y_{it} &= p + ry_{i,t-1} + u_{it}, \\ t &= 1,...,T_i, \quad i = 1,...,n, \end{aligned} \tag{4}$$

so the corresponding estimators are easily obtained with common econometric packages. For $n \to \infty$, these estimators are consistent, $\sqrt{n}$-asymptotically normal with the usual covariance matrix associated with (heteroskedastic) NLS estimation.

The CLS procedure assumes homoskedastic errors, which is not so by construction. Under correct specification of $VAR\left(y_{it}|y_{i,t-1}\right)$ an alternative estimator, more efficient than CLS is conditional weighted least squares (CWLS), with the weighting functions obtained from

$$\begin{aligned} VAR\left(u_{it}|y_{i,t-1}\right) &= VAR\left(y_{it}|y_{i,t-1}\right) \\ &= y_{i,t-1}^2 p\left(1-p\right)\left(1-p_1\right)^2 + y_{i,t-1}\left(1-p\right)\left(1-p_1\right)\left(2p+p_1\right) + p\left(1-p\right), \end{aligned} \tag{5}$$

replacing p and p_1 with its CLS estimates. As the true form of heteroskedasticity may be unknown, the robust "sandwich" covariance estimator should also be used for inference purposes (either with CLS or CWLS).

As previously mentioned, the assumption of constant parameters, p and p_1, $\forall i$, can be rather restrictive. This assumption can be relaxed, even within a pure time series context.[5] Within a regression framework one can specify p_i

[5] The assumption can be relaxed within a pure time series context by adopting individual-specific parameters, p_i and p_{1i}. Then,

$$\begin{aligned} y_{it} &= p_i + r_i y_{i,t-1} + u_{it}, \\ u_{it} &= \left(d_{it}-p_i\right)\left(y_{i,t-1}+1\right) + \left(1-d_{it}\right)\left(p_{1i}\circ y_{i,t-1}\right) + \left(1-p_i\right)\left(p_{1i}y_{i,t-1}\right), \end{aligned}$$

where $E\left(d_{it}|p_i\right) = p_i$, $r_i \equiv p_i + p_{1i}\left(1-p_i\right)$ and $p_{1i}\circ y_{i,t-1} \equiv \sum_{j=1}^{y_{i,t-1}} b_{ij}$, with $E\left(b_{ij}|p_{1i}\right) = p_{1i}$. The model parameters can be estimated through separate CLS regressions (one for each individual with $T_i > 2$), or a single regression with individual dummies – the resulting estimators are consistent with $T_i \to \infty$. The precision of individual estimates of this fixed

and p_{1i}, for instance, as in (2). Accordingly, the CLS estimator of $\theta \equiv \left(\beta', \gamma'\right)'$ can be defined as

$$\hat{\theta}_{CLS} = \arg\min_{\theta} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left(y_{it} - p_i - r_i y_{i,t-1}\right)^2.$$

This estimator is also relatively straightforward to implement, given access to a software package including NLS estimation, although reported standard errors may wrongly assume homoskedastic errors. CWLS is a more efficient method, for which the weighting functions can be obtained from (3) and (5), with expectations now conditional on x_i and $y_{i,t-1}$. Again, if the true form of heteroskedasticity is unknown, the robust "sandwich" covariance matrix estimator should also be used for inference purposes.

Accounting for unobserved individual heterogeneity can be difficult in a regression context if the form of this heterogeneity is left unspecified. A feasible approach introduces unobservables in the model only through p_i and p_{1i}. One presumably useful simplification considers individual-specific, time-invariant random intercepts in both p_i and p_{1i}, leading to the formal definitions[6]

$$\begin{aligned} p_i &\equiv \left(1 + \exp\left(-x_i'\beta - \varepsilon_i\right)\right)^{-1}, \\ p_{1i} &\equiv \left(1 + \exp\left(-z_i'\gamma - \varepsilon_i\right)\right)^{-1}. \end{aligned} \tag{6}$$

Then, for each individual the model for the conditional first moment of y_{it} becomes

$$\begin{aligned} E\left(y_{it}|y_{i,t-1}, x_i\right) &= q_i + r_i y_{t-1}, \\ q_i &\equiv \int p_i f_\varepsilon\left(\varepsilon\right) d\varepsilon, \\ r_i &\equiv \int \left(p_i + p_{1i}\left(1 - p_i\right)\right) f_\varepsilon\left(\varepsilon\right) d\varepsilon \end{aligned} \tag{7}$$

where $\varepsilon \equiv \varepsilon_i$, $i = 1, ..., n$, denote i.i.d. random variables with zero mean and some specified density f_ε. It is noted that, conditional on observed regressors,

effects approach can be poor for individuals with low T_i. Still, it can be useful as a first step in CWLS estimation or as a means to test the hypothesis of parameter constancy – with, *e.g.*, a standard F test of model (4) against the latter specification. Alternatively, a random effects approach can be adopted, with estimation of the relevant parameters of the distributions of the random variables p_i and p_{1i} over different individuals.

[6] Obviously, unobserved individual heterogeneity can be dealt with in this same way within a pure time series approach – the only difference being the absence of *observable* regressors.

p_i and p_{1i} are no longer independent.[7] The parameters of the model (β, γ and parameters in f_ε) can be estimated through nonlinear CLS, provided intervening integrals are conveniently computed (with quadrature or simulation approximation methods, if necessary).[8]

Accounting for unobservables in this panel data regression model has now led to full specification of the conditional distribution of the sequence $\{y_t, t = 1, ..., T\}$ given x. Though easy to implement, the CLS approach does not make use of all the information this full specification provides. One obvious alternative, possibly more efficient, is provided by the ML method, to be addressed in the next section.

3.2 Maximum Likelihood

Under the assumption of independent individual time sequences ML estimation requires full specification of the joint probabilistic model for the Markov chain $\{y_t, t = 1, ..., T\}$. This model is already implicit in previous assumptions – (1)

[7] In general, accounting for unobserved individual heterogeneity affects the Markov nature of each individual sequence $\{y_t, t = 1, ..., T\}$. For simplicity, let $\varepsilon_{it} = \varepsilon_i$, $\forall t$. If, conditional on x and ε, the sequence constitutes a Markov chain, that is,

$$f(y_1, ..., y_T|\varepsilon) = f(y_1|\varepsilon) \prod_{t=2}^{T} f(y_t|y_{t-1}, \varepsilon)$$

(conditioning on x, and function indices omitted), the same is not true when ε is integrated out from this joint density. Formally,

$$\begin{aligned} f(y_1, ..., y_T) &= \int f(y_1|\varepsilon) \prod_{t=2}^{T} f(y_t|y_{t-1}, \varepsilon) f_\varepsilon(\varepsilon) d\varepsilon \\ &\neq \int f(y_1|\varepsilon) f_\varepsilon(\varepsilon) d\varepsilon \prod_{t=2}^{T} \int f(y_t|y_{t-1}, \varepsilon) f_\varepsilon(\varepsilon) d\varepsilon \\ &= f(y_1) \prod_{t=2}^{T} f(y_t|y_{t-1}). \end{aligned}$$

Nevertheless, the formal inclusion of ε in the model (through p and p_1) causes y_t to be dependent on other elements of the chain only through y_{t-1} and ε. This leads to consideration of the conditional first moment of y_t as in (7), thereby enabling CLS estimation analogously as before.

[8] Other schemes could be adopted for the inclusion of unobservables, such as random β and/or γ coefficients (other than the intercepts), time-varying individual effects and/or different (correlated) disturbances in p and p_1. Presumably, the usefulness of such sophistications should be weighted against the added estimation difficulty they are bound to represent.

and (2) (or (7) with specification of f_ε). Here, this full specification is made explicit in order to implement ML.

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) leads to the following model of transition probabilities (individual indices omitted, so $T \equiv T_i$):

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

where $\mathbf{1}(\cdot)$ denotes the usual indicator function. Then, the joint conditional density of each individual sequence $\{y_t, t=1,...,T\}$ can be written as

$$\begin{aligned}
& f_y(y_1,...,y_T|p,p_1) \\
= & \prod_{t=1}^{T}\left(p^{\mathbf{1}(y_t=y_{t-1}+1)}\left((1-p)\binom{y_{t-1}}{y_t}p_1^{y_t}(1-p_1)^{y_{t-1}-y_t}\right)^{\mathbf{1}(y_t\leq y_{t-1})}\right) \\
= & \prod_{t=1}^{T}\left(\left(\frac{p}{1-p}\right)^{\mathbf{1}(y_t=y_{t-1}+1)}(1-p)\right. \\
& \left.\times\left(\binom{y_{t-1}}{y_t}\left(\frac{p_1}{1-p_1}\right)^{y_t}(1-p_1)^{y_{t-1}}\right)^{\mathbf{1}(y_t\leq y_{t-1})}\right) \\
\propto & \left(\frac{p}{1-p}\right)^{\sum_{t=1}^{T}\mathbf{1}(y_t=y_{t-1}+1)}(1-p)^T \\
& \times\left(\left(\frac{p_1}{1-p_1}\right)^{\sum_{t=2}^{T}\mathbf{1}(y_t\leq y_{t-1})y_t}(1-p_1)^{\sum_{t=2}^{T}\mathbf{1}(y_t\leq y_{t-1})y_{t-1}}\right)^{\mathbf{1}(T\geq 2)},
\end{aligned} \tag{8}$$

exhibiting, as expected, the usual split of hurdle models into 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,

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

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 $y_{it} = y_{i,t-1} + 1$.

If one wants to analyze a regression model, covariates can be introduced through parameters p and p_1, specified, for instance, as logits.[9] Unobserved individual heterogeneity can also be dealt with, analogously as in the previous section. However, if such heterogeneity is present in both p and p_1, ML estimation is bound to become more difficult because, in general, the likelihood no longer factors into two functionally independent terms.[10] In addition, as the likelihood may involve complex integrals with no known analytical expressions, simulation or quadrature approximation techniques may be required to estimate the model.

4 Simulation Results

4.1 Monte Carlo Design

This section illustrates the performance of the CLS and ML estimators of the panel data hurdle model on the basis of a simulated panel data set. This panel contains information on each individual time series of counts up to the sampling date, as well as on a set of covariates.

As before, 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). Two samples were generated: the first with $n = 360$ individual time sequences and the second with $n = 5400$ sequences. In each sample $1 \leq T_i \leq 36$, with the same number of individuals for each different T_i value. That is, the

[9] The above expressions hold with time-invariant regressors. If time-varying covariates are introduced, notation must be altered accordingly.

[10] This is in line with the approach of Winkelmann (2004). In this case the individual contribution to the likelihood can be written as

$$L = \int \int A(p)\, B(p_1)\, dF(p, p_1),$$

where A and B denote, respectively, the first and second parts of the hurdle likelihood, and $F(p, p_1)$ denotes the joint distribution of p and p_1 (possible conditioning on observed regressors omitted). Only with independent p and p_1 can this expression be written as a product of two functionally independent terms.

smaller sample contains 10 (= 360/36) individuals per each T_i, whereas the larger sample contains 150 (= 5400/36) individuals per each T_i.

Individual time sequences are drawn independently (independence across i); each individual sequence $\{y_{it}, t = 1, ..., T_i\}$ is drawn from the joint (conditional) p.f. defined in (8) with p and p_1 specified as logits. These include the following regressors from the corresponding marginals: $x_1 \equiv 1$ (intercept term), $x_2 \sim$ *Bernoulli* (.25), $x_3 \sim N(0, 2.25)$, and x_4 a rank variable generated by

$$x_4 = \begin{cases} 1, & 0 < w < 2, \\ 2, & 2 \leq w < 4, \\ 3, & 4 \leq w < 6, \\ 4, & 6 \leq w < 8, \\ 5, & w \geq 8, \end{cases}$$

with w a chi-squared variable with three degrees of freedom. Covariates are assumed time-invariant ($x_{kit} \equiv x_{ki}$, $k = 2, 3, 4$) with p and p_1 specified as

$$\begin{aligned} p_{it} &\equiv (1 + \exp(-x_i'\beta - \lambda y_{i,t-1} - \varepsilon_i))^{-1}, \\ p_{1i} &\equiv (1 + \exp(-x_i'\gamma - \varepsilon_i))^{-1}, \end{aligned}$$

with $x_i \equiv (1, x_{2i}, x_{3i}, x_{4i})'$, $\beta \equiv (.85, -.75, .2, -1.5)'$, $\gamma \equiv (1, -.5, .5, -.5)'$, λ is a parameter and $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$, $\forall i$, denotes an unobserved individual effect. For each of the two sample sizes various DGP's are considered, according to different values of λ and σ_ε^2. Respectively, $\lambda \in \{0, -.3\}$ and $\sigma_\varepsilon^2 \in \{0, 1\}$. For instance, $\left(\lambda, \sigma_\varepsilon^2\right) = (0, 0)$ means that the sample is generated without unobserved individual heterogeneity and with the assumption that, at each date, individual decisions are independent of the previous value of the count variable ($y_{i,t-1}$). Alternatively a negative value is chosen for λ, under which p_{it} is lower the higher $y_{i,t-1}$. The following table labels these DGP's:

Data Generating Processes

	DGP1	DGP2	DGP3	DGP4
λ	0	$-.3$	0	$-.3$
σ_ε^2	0	0	1	1

These simulated data sets can be thought of as samples of loan repayment histories in fixed installments. Each individual represents a different contract aged T_i months at the sampling date. The assumption of time-invariant regressors may reflect the frequent fact that contracts and clients characteristics are recorded at the time of loan applications and remain subsequently unchanged over the repayment period. A negative value for λ (p_{it} decreasing in $y_{i,t-1}$) is naturally interpreted as a decrease in the probability *not* to pay, in response to an increase in the number of previously missed installments.

In the simulation experiment model (1) is denoted, respectively, as PHL and PH, according to whether it is specified with or without lagged dependent variable in p. The performance of CLS and ML estimation methods is assessed on the basis of 2000 replications of the described samples with regressors newly drawn at each replica. All computations were performed using TSP 5.0 (Hall and Cummins, 2005).

4.2 Monte Carlo Results

Estimation results for models PH and PHL are displayed in tables 1.1 through 4.2, corresponding to the four DGP's and two different sample sizes. The tables contain estimates averages and standard errors for parameters in p (β_1 through β_4, and λ for model PHL) and in p_1 (γ_1 through γ_4). ML estimates and standard errors for the γ parameters are the same under models PH and PHL because the second part of the log-likelihood (involving p_1) is identical in both models (see (8)). Obviously, this is not the case for the CLS estimates of γ.

The estimation results suggest some remarks. First, under DGP1 both models are correctly specified. Accordingly, β estimates from both models are relatively similar. Both CLS and ML seem to work slightly better (with lower standard errors) under model PH than with PHL, which is expectable because the former assumes the true value for λ, while it is estimated in the latter. For CLS this finding extends to p_1 parameters as well. Understandably, the advantage of model PH over PHL is reduced as the sample size increases. Meanwhile, under DGP1 ML standard errors are invariably lower than CLS, a consequence of the former method full usage of the available information.

Under DGP2 only model PHL is correctly specified. Nevertheless, model PH seems to produce reasonable CLS β estimates with both sample sizes, quite

better than ML and close to PHL β estimates. Actually this pattern can be deceiveing as it appears to be associated with a relatively low absolute value of λ, bringing models PH and PHL close to one another. Unreported simulation results for model PH do indicate a strong sensitivity of both β and γ CLS estimates to the true value of λ. Meanwhile, for the reason already mentioned, ML estimates of γ coincide with those from model PHL and are, therefore, obtained from consistent estimators. Given the functional separability of the likelihood function, ML γ estimates are immune to the misspecification of p.

Under DGP3 and DGP4 no model is correctly specified, as none of them allows for unobserved individual heterogeneity. The consequences of such misspecification appear somewhat widespread, affecting both p and p_1 parameters estimates (tables 3.1 through 4.2). These effects are particularly severe for the CLS estimates of the intercept term in p_1 but, overall, results are not trustworthy under neglected individual heterogeneity.

5 Conclusion

This paper presents a hurdle model for panel count data with excess zeros and bounded support, suggesting its estimation through CLS and ML. The use and behaviour of both methods is illustrated with simulated data sets of independent individual time sequences of counts.

The performance of CLS and ML appears somewhat similar under the considered DGP's, with a slight and expectable advantage of the latter in cases of correct model specification. Naturally, efficiency advantages of ML should be weighted against the considerably easier coding of CLS estimation.

The proposed specification can be used, for instance, as a model of repayment behaviour to be employed in a context of credit or behavioural scoring. In this sense it may provide better estimates of default probability than, *e.g.*, simple cross-section models of the total number of missed installments. Meanwhile, with some adjustments, the approach is rich enough to encompass such situations as early repayment or redemption of loans, as well as loan contracts with variable installments. Naturally, these examples can provide the ground for future work.

Estimation Results - DGP1

Table 1.1 - Sample: 360 individuals

Model	PH				PHL			
Est. Meth.	CLS		ML		CLS		ML	
	est.	s.d.	est.	s.d.	est.	s.d.	est.	s.d.
β_1	.876	.184	.851	.096	.902	.180	.790	.112
β_2	-.753	.150	-.752	.093	-.764	.151	-.732	.094
β_3	.207	.047	.201	.024	.213	.049	.190	.026
β_4	-1.513	.115	-1.502	.067	-1.531	.117	-1.477	.070
λ	–		–		-.051	.103	.028	.026
γ_1	.949	.338	1.004	.129	1.161	.555	1.004	.129
γ_2	-.519	.291	-.506	.128	-.578	.322	-.506	.128
γ_3	.497	.098	.501	.034	.523	.113	.501	.034
γ_4	-.488	.191	-.507	.096	-.575	.262	-.507	.096

Table 1.2 - Sample: 5400 individuals

Model	PH				PHL			
Est. Meth.	CLS		ML		CLS		ML	
	est.	s.d.	est.	s.d.	est.	s.d.	est.	s.d.
β_1	.852	.047	.849	.025	.854	.045	.779	.029
β_2	-.751	.038	-.751	.024	-.752	.038	-.728	.024
β_3	.200	.012	.200	.006	.201	.012	.188	.007
β_4	-1.501	.029	-1.499	.018	-1.502	.029	-1.471	.018
λ	–		–		-.003	.021	.032	.006
γ_1	.995	.085	1.000	.033	1.011	.142	1.000	.033
γ_2	-.501	.076	-.500	.032	-.505	.082	-.500	.032
γ_3	.501	.024	.500	.009	.503	.029	.500	.009
γ_4	-.499	.048	-.500	.025	-.505	.067	-.500	.025

Estimation Results - DGP2

Table 2.1 - Sample: 360 individuals

Model	PH				PHL			
Est. Meth.	CLS		ML		CLS		ML	
	est.	s.d.	est.	s.d.	est.	s.d.	est.	s.d.
β_1	.838	.152	.433	.087	.899	.156	.742	.114
β_2	-.741	.133	-.629	.090	-.766	.140	-.721	.099
β_3	.199	.036	.133	.023	.208	.040	.183	.027
β_4	-1.468	.098	-1.341	.062	-1.521	.104	-1.459	.071
λ	–		–		-.298	.184	-.214	.042
γ_1	-.238	.299	1.005	.137	.842	.588	1.005	.137
γ_2	-.199	.259	-.506	.134	-.476	.273	-.506	.134
γ_3	.413	.074	.503	.036	.495	.085	.503	.036
γ_4	-.024	.163	-.507	.102	-.444	.257	-.507	.102

Table 2.2 - Sample: 5400 individuals

Model	PH				PHL			
Est. Meth.	CLS		ML		CLS		ML	
	est.	s.d.	est.	s.d.	est.	s.d.	est.	s.d.
β_1	.817	.037	.431	.022	.853	.041	.733	.028
β_2	-.734	.033	-.624	.022	-.750	.035	-.714	.023
β_3	.195	.009	.132	.006	.200	.010	.181	.007
β_4	-1.456	.024	-1.340	.016	-1.501	.027	-1.454	.018
λ	–		–		-.301	.061	-.209	.010
γ_1	-.190	.073	1.001	.035	.989	.176	1.001	.035
γ_2	-.188	.066	-.500	.034	-.498	.074	-.500	.034
γ_3	.412	.018	.500	.009	.500	.022	.500	.009
γ_4	-.037	.040	-.500	.026	-.496	.073	-.500	.026

Estimation Results - DGP3

Table 3.1 - Sample: 360 individuals

Model	PH				PHL			
Est. Meth.	CLS		ML		CLS		ML	
	est.	s.d.	est.	s.d.	est.	s.d.	est.	s.d.
β_1	-.359	.362	.798	.170	-.344	.335	.149	.147
β_2	-.546	.363	-.658	.177	-.451	.306	-.499	.133
β_3	.212	.124	.174	.048	.064	.100	.091	.035
β_4	-1.189	.264	-1.330	.101	-1.009	.216	-1.108	.083
λ	–		–		.179	.190	.214	.024
γ_1	7.011	5.059	1.134	.209	2.306	1.999	1.134	.209
γ_2	-1.730	4.297	-.282	.219	-.350	.939	-.282	.219
γ_3	1.184	.996	.386	.059	.516	.468	.386	.059
γ_4	-1.546	1.571	-.144	.142	-.202	.832	-.144	.142

Table 3.2 - Sample: 5400 individuals

Model	PH				PHL			
Est. Meth.	CLS		ML		CLS		ML	
	est.	s.d.	est.	s.d.	est.	s.d.	est.	s.d.
β_1	-.443	.102	.790	.044	-.407	.081	.137	.038
β_2	-.635	.117	-.645	.046	-.411	.074	-.487	.034
β_3	.263	.032	.172	.012	.056	.022	.089	.009
β_4	-1.184	.074	-1.323	.026	-.904	.046	-1.101	.021
λ	–		–		.235	.014	.215	.006
γ_1	6.184	.544	1.123	.054	1.464	.257	1.123	.054
γ_2	-1.070	.529	-.260	.056	-.094	.155	-.260	.056
γ_3	.803	.143	.381	.015	.299	.051	.381	.015
γ_4	-1.444	.247	-.130	.036	.061	.130	-.130	.036

Estimation Results - DGP4

Table 4.1 - Sample: 360 individuals

Model	PH				PHL			
Est. Meth.	CLS		ML		CLS		ML	
	est.	s.d.	est.	s.d.	est.	s.d.	est.	s.d.
β_1	-.415	.280	.266	.125	-.164	.259	.079	.142
β_2	-.424	.281	-.534	.139	-.463	.246	-.487	.131
β_3	.059	.090	.102	.036	.047	.080	.078	.035
β_4	-1.017	.172	-1.170	.080	-1.041	.159	-1.107	.081
λ	–		–		.120	.436	.095	.030
γ_1	2.536	.698	1.309	.231	.930	1.123	1.309	.231
γ_2	-.506	.659	-.347	.240	-.125	.549	-.347	.240
γ_3	.504	.232	.406	.065	.351	.216	.406	.065
γ_4	-.451	.417	-.228	.148	.109	.478	-.228	.148

Table 4.2 - Sample: 5400 individuals

Model	PH				PHL			
Est. Meth.	CLS		ML		CLS		ML	
	est.	s.d.	est.	s.d.	est.	s.d.	est.	s.d.
β_1	-.512	.075	.255	.031	-.262	.061	.062	.036
β_2	-.434	.079	-.528	.036	-.421	.057	-.480	.033
β_3	.079	.023	.102	.009	.041	.017	.076	.009
β_4	-.963	.045	-1.160	.020	-.958	.037	-1.096	.020
λ	–		–		.251	.013	.097	.008
γ_1	2.598	.167	1.300	.060	.557	.166	1.300	.060
γ_2	-.424	.171	-.322	.061	.006	.108	-.322	.061
γ_3	.426	.054	.404	.017	.270	.035	.404	.017
γ_4	-.465	.102	-.216	.039	.246	.091	-.216	.039

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 CLS estimator, with $n \to \infty$.

Section 2 — $E(y_t) = p\left(1 + r + ... + r^{t-1}\right)$.

Proof. Under the definition of the model in (1) and subsequent assumptions, $E(y_1) = E(d_1) = p$. That is, the proposed formula is valid for $t = 1$. Supposing that $E(y_t) = p\left(1 + r + ... + r^{t-1}\right)$, it follows that

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

which confirms the generality of the proposed result, for any positive integer t. ■

Section 2 — $COV(y_t, y_{t-k}) = r^k VAR(y_{t-k})$.

Proof. *i.* The formula is derived on the basis of the following preliminary result for the conditional first moment of y_t, given y_{t-k}:

$$E(y_t|y_{t-k}) = p\left(1 + r + ... + r^{k-1}\right) + r^k y_{t-k}. \tag{9}$$

This can be shown by mathematical induction as well. It is first noted that, given the Markov nature of the sequence $\{y_t\}$, one can write

$$E(y_t|y_{t-k-1}) = E\left(E(y_t|y_{t-k})|y_{t-k-1}\right). \tag{10}$$

Now, for $k = 1$, (9) is verified, because $E(y_t|y_{t-1}) = p + ry_{t-1}$. Then, (9) and (10) lead to the required result

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

ii. From

$$\begin{aligned} & COV\left(y_t, y_{t-k}\right) \\ = \; & E\left(y_t\left(y_{t-k} - E\left(y_{t-k}\right)\right)\right) = E\left(E\left(y_t|y_{t-k}\right)\left(y_{t-k} - E\left(y_{t-k}\right)\right)\right), \end{aligned}$$

and from *i.*, it follows that

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

■

Section 3.1 — Consistency and $\sqrt{n}$-asymptotic normality of the CLS estimator of $(p, p_1)'$ (bounded T_i, $\forall i$).

Proof. Asymptotic properties are stated with respect to $(p, r)'$ for the case $n \to \infty$. From these the desired properties of the CLS estimator of $(p, p_1)'$ follow immediately, with the asymptotic covariance matrix of the $(p, p_1)'$ estimator obtained through the delta method.

Let $\pi \equiv (p, r)'$ and $\hat{\pi}$ denote its CLS estimator. Then, one can write

$$\hat{\pi} = \pi + \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} z_{it} z_{it}'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} z_{it} u_{it}\right),$$

with $z_{it} \equiv (1, y_{i,t-1})'$, $y_{i0} \equiv 0$ and $u_{it} \equiv y_{it} - p - p_1 y_{i,t-1}$. Suppose, for simplicity, that $p_1 < 1$ (the case $p_1 = 1$ does not affect asymptotic properties for fixed T_i, $\forall i$). With fixed T_i, $\sum_{t=1}^{T_i} z_{it} z_{it}'$ is necessarily bounded, so $\sum_{i=1}^{n}\sum_{t=1}^{T_i} z_{it} z_{it}'$ is $O_p(n)$. Also, $p\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} z_{it} u_{it} = 0$, because of the following readily established results: $E\left(u_{it}\right) = 0$, $E\left(u_{it} y_{i,t-1}\right) = 0$, and

$$\begin{aligned} & \lim_{n\to\infty}\begin{bmatrix} VAR\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} u_{it}\right) \\ VAR\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} y_{i,t-1} u_{it}\right) \end{bmatrix} \\ = \; & \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\begin{bmatrix} \sum_{t=1}^{T_i} VAR\left(u_{it}\right) \\ \sum_{t=1}^{T_i} VAR\left(y_{i,t-1} u_{it}\right) \end{bmatrix}, \end{aligned}$$

because cross-product terms are null in $VAR\left(\sum_{t=1}^{T_i} z_{it} u_{it}\right)$, given the fact that

$$VAR\left(\sum_{t=1}^{T_i} u_{it}\right) = \sum_{t=1}^{T_i} VAR\left(u_{it}\right)$$

and

$$VAR\left(\sum_{t=1}^{T_i} y_{i,t-1}u_{it}\right) = \sum_{t=1}^{T_i} VAR\left(y_{i,t-1}u_{it}\right),$$

as y_{it} and $y_{i,t-k}$ are independent, conditionally on $y_{i,t-1}$. Then, the last expression is equal to

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \begin{bmatrix} T_i\left(p\left(1-p\right)+pp_1\right) - \left(1-p_1^{T_i}\right)/\left(1-p_1\right) \\ p\left(1-p\right)E\left(y_{i,t-1}^2\right) + p_1\left(1-p_1\right)E\left(y_{i,t-1}^3\right) \end{bmatrix} = 0,$$

because $E\left(y_{it}^3\right)$ is necessarily bounded for fixed T_i.[11]

Therefore, the CLS estimator of $(p,r)'$ is consistent. Asymptotic normality of $\sqrt{n}\left(\hat{\pi}-\pi\right)$ follows immediately from the fact that both elements of $\sum_{t=1}^{T_i} z_{it}u_{it}$ ($\sum_{t=1}^{T_i} u_{it}$ and $\sum_{t=1}^{T_i} y_{i,t-1}u_{it}$) have zero mean and bounded variance for bounded T_i, enabling application of a central limit theorem to the sequence $\left\{\sum_{t=1}^{T_i} z_{it}u_{it}\right\}$. In matrix notation one can write

$$\sqrt{n}\left(\hat{\pi}-\pi\right) = \left(\frac{1}{n}Z'Z\right)^{-1}\frac{1}{\sqrt{n}}Z'u,$$

where

$$\underset{\left(\sum T_i \times 2\right)}{Z} \equiv \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}, \quad \underset{\left(T_i\times 2\right)}{Z_i} \equiv \begin{bmatrix} 1 & 0 \\ 1 & y_{i1} \\ \dots & \dots \\ 1 & y_{i,T_i-1} \end{bmatrix},$$

and u denotes the column $\left(\sum T_i\right)-$ vector of residuals, $y_{it}-\left(p+ry_{i,t-1}\right)$, $t=1,...,T_i$, $i=1,...,n$. Thus,

$$\sqrt{n}\left(\hat{\pi}-\pi\right) \dot{\sim} N\left(0,V\right),$$

[11] Specifically,

$$E\left(y_{it}^2\right) = p\frac{1-p_1^t}{1-p_1} - p^2\frac{1-p_1^{2t}}{1-p_1^2} + p^2\left(\frac{1-p_1^t}{1-p_1}\right)^2,$$

and

$$E\left(y_{it}^3|y_{i,t-1}\right) = p_1^3y_{i,t-1}^3 + \left(3p_1^2p - 3p_1^3 + 3p^2\right)y_{i,t-1} + \left(6p_1p - 3p_1^2p - 3p_1^2 + 2p_1^3 + p_1\right)y_{i,t-1} + p,$$

as can be obtained, for instance, from the conditional moment generating function of y_t, given y_{t-1},

$$E\left(\exp\left(sy_t\right)|y_{t-1}\right) = \left(1-p+pe^s\right)\left(1-p_1+p_1e^s\right)^{y_{t-1}},$$

(the expression for the unconditional moment, $E\left(y_{it}^3\right)$ is quite cumbersome to obtain).

where

$$V \equiv \left(p \lim_{n\to\infty} \frac{1}{n} Z'Z \right)^{-1} \left(p \lim_{n\to\infty} \frac{1}{n} Z'uu'Z \right) \left(p \lim_{n\to\infty} \frac{1}{n} Z'Z \right)^{-1}.$$

(The formal expression for this variance involves the unconditional variances of y_{it}, $t = 1, ..., T_i$). ■

Section 3.1 — CWLS estimation, expression (5):

$$\begin{aligned} & VAR\left(y_{it}|y_{i,t-1}\right) \\ = \; & y_{i,t-1}^2 p\left(1-p\right)\left(1-p_1\right)^2 + y_{i,t-1}\left(1-p\right)\left(1-p_1\right)\left(2p+p_1\right) + p\left(1-p\right). \end{aligned}$$

Proof. The model assumptions lead to

$$\begin{aligned} E\left(y_t^2|y_{t-1}\right) &= pE\left(y_t^2|y_{t-1}, d_t = 1\right) + \left(1-p\right) E\left(y_t^2|y_{t-1}, d_t = 0\right) \\ &= p\left(y_{t-1}+1\right)^2 + \left(1-p\right)\left(V\left(p_1 \circ y_{t-1}\right) + E^2\left(p_1 \circ y_{t-1}\right)\right) \\ &= p\left(y_{t-1}+1\right)^2 + \left(1-p\right)\left(y_{t-1}p_1\left(1-p_1\right) + y_{t-1}^2 p_1^2\right). \end{aligned}$$

Then, expression (5) follows from the equality

$$VAR\left(y_t|y_{t-1}\right) = E\left(y_t^2|y_{t-1}\right) - E^2\left(y_t|y_{t-1}\right),$$

with $E\left(y_t|y_{t-1}\right) = p + \left(p + p_1\left(1-p\right)\right) y_{t-1}$. ■

References

- Al-Osh, M. A., A. A. Alzaid (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, no. 355, University of Umeå.
- Brännäs, K. (1995), "Explanatory Variables in the AR(1) Model", Umeå Economic Studies, no. 381, University of Umeå.
- Cameron, A. C., P. K. Trivedi (1998), *Regression Analysis of Count Data*, Cambridge, Cambridge University Press.
- Cragg, J. G. (1971), "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods", *Econometrica*, 39, 829-844.
- Hall, B. H. and Cummins, C. (2005). *TSP 5.0 User's Guide*, TSP International, Palo Alto (CA).
- Jacobs, P. A., P. A. W. Lewis (1978), "Discrete Time Series Generated by Mixtures II: Asymptotic Properties," *Journal of The Royal Statistical Society B*, 40, 222-228.
- Jin-Guan, D., L. Yuan (1991), "The Integer-Valued Autoregressive (INAR(p)) Model", *Journal of Time Series Analysis*, 12, 129-142.
- 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 D. N. Shanbhag, C. R. Rao, eds., *Handbook of Statistics*, vol. 21, pp. 573-606, Amsterdam, North-Holland.
- Mullahy, J. (1986), "Specification and Testing of Some Modified Count Data Models", *Journal of Econometrics*, 33, 341-365.

- Ronning, G., R. C. Jung (1992), "Estimation of a First Order Autoregressive Process with Poisson Marginals for Count Data", in L. Fahrmeir, et al., eds., *Advances in GLIM and Statistical Modelling*, New York, Springer.
- Steutel, F., K. VanHarn (1979), "Discrete Analogues of Self-Decomposability and Stability", *Annals of Probability*, 7, 893-899.
- Thomas, L. C., D. B. Edelman, J. N. Crook (2002), *Credit Scoring and its Applications,* Philadelphia, SIAM.
- Windmeijer, F. (2006), "GMM for Panel Count Data Models", Discussion Paper No. 06/591, University of Bristol, Dept. of Economics.
- Winkelmann, R. (2004), "Health Care Reform and the Number of Doctor Visits - An Econometric Analysis", *Journal of Applied Econometrics*, 19(4), 455-472.