Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Journal of Econometrics 141 (2007) 1073–1099

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

A zero-inflated ordered probit model, with an application to modelling tobacco consumption

Mark N. Harris*, Xueyan Zhao

Department of Econometrics and Business Statistics, Building 11, Wellington Road, Clayton Campus, Monash University, Vic. 3800, Australia

Available online 26 March 2007

Abstract

Data for discrete ordered dependent variables are often characterised by "excessive" zero observations which may relate to two distinct data generating processes. Traditional ordered probit models have limited capacity in explaining this preponderance of zero observations. We propose a zero-inflated ordered probit model using a double-hurdle combination of a split probit model and an ordered probit model. Monte Carlo results show favourable performance in finite samples. The model is applied to a consumer choice problem of tobacco consumption indicating that policy recommendations could be misleading if the splitting process is ignored.

JEL classification: C3; D1; I1

Keywords: Ordered outcomes; Discrete data; Tobacco consumption; Zero-inflated responses

1. Introduction and Background

Often in empirical economics interest lies in modelling a discrete random variable that is inherently ordered. Examples include survey responses on opinions, employment status levels, bond ratings and job classifications by skill levels. Typically, the empirical strategy employed would involve estimation of an ordered probit (OP) or logit model (see, for example, McKelvey and Zavoina, 1975; Marcus and Greene, 1985). However, often data

*Corresponding author. Tel.: +61 39905 9414; fax: +61 39905 5474.

E-mail addresses: mark.harris@buseco.monash.edu.au (M.N. Harris), xueyan.zhao@buseco.monash.edu.au (X. Zhao).

0304-4076/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2007.01.002

for such ordered random variables are characterised by excessive observations in the choice at one end of the ordering or, typically, zeros. For example, in a survey corresponding to illicit drug use, answers to a question such as "*how often do you use drug X?*" are likely to be characterised by an excess of zero observations when discrete choices of consumption levels including "*never/not recently*" ($y = 0$) are presented.

Traditional OP models have limited capacity in explaining such a preponderance of zero observations, especially when the zeros may relate to two distinct sources. In the case of discrete levels of reported drug consumption, zeros will be recorded for non-participants who, for example, abstain due to health or legal concerns and who pay no regard to drug in their decision making. However, there may also be zeros who are the corner solution of a standard consumer demand problem and who may become consumers if the price is lower or income is higher. Thus, it is likely that these two types of zeros are driven by different systems of consumer behaviour. Here, zero consumption potential users are likely to possess characteristics similar to those of the users and are likely to be responsive to standard consumer demand factors such as prices and income. On the other hand, genuine non-participants are likely to have perfectly inelastic price and income demand schedules, and are driven by a separate process relating to sociological, health and ethical considerations. If such underlying processes are modelled incorrectly, it could invalidate any subsequent policy implications. Additionally, even the same explanatory variable could have different effects on the two decisions. One example is the effect of income on drug consumption. Higher income, acting as an indicator for social class and health awareness, may increase the chance of genuine non-participation. However, for participants, higher income will be associated with lower chances of zero consumption, if tobacco is a normal good, for these participants. An OP model generated by a single latent equation cannot allow for the differentiation between the two opposing effects.

In a manner analogous to the zero-inflated/augmented Poisson (ZIP/ZAP) models in the count data literature (see, for example, Mullahey, 1986; Heilbron, 1989; Lambert, 1992; Greene, 1994; Pohlmeier and Ulrich, 1995; Mullahey, 1997) and double-hurdle models in the limited dependent variable literature (see, for example, Cragg, 1971), this paper proposes an extension to the OP model to take into account of the possibility that the zeros can arise from two different aspects of individual behaviour. Unlike the Poisson and negative binomial regression framework, the ultimate data generating process here can be seen as coming from two separate underlying latent variables. We propose a zero-inflated ordered probit (ZIOP) model that involves a system of a probit "splitting" model and an OP model which relate to potentially differing sets of covariates. We also further allow the error terms of the two latent equations to be correlated (denoted a ZIOPC model), along the lines of a Heckman-selection-type model (Heckman, 1979).

Monte Carlo experiments are conducted under various true models to examine the finite sample performance. We also report performances of various specification tests and model selection criteria for choosing between the OP, ZIOP and ZIOPC models. The model is then applied to a unit record dataset from Australia on tobacco consumption, which involves an estimation sample of nearly 29,000 individuals and 76% of zero observations. The application clearly illustrates the extra insights provided by the ZIOP/ZIOPC model in analysing the effects of some important explanatory factors on individuals' tobacco consumption patterns.

2. The econometric framework

2.1. A zero-inflated ordered probit (ZIOP) model

We start by defining a discrete random variable y that is observable and assumes the discrete ordered values of $0, 1, \ldots, J$. A standard OP approach would map a single latent variable to the observed outcome y via so-called boundary parameters, with the latent variable being related to a set of covariates. Here we propose a ZIOP model that involves two latent equations: a probit selection equation and an OP equation. This splits the observations into two regimes that relate to potentially two different sets of explanatory variables. Consider the drug consumption example. Here an individual user is modelled as having to overcome two hurdles: whether to participate, and then, conditional on participation, how much to consume *which also includes zero consumption*. The two types of zero-consumption observations relate to those non-participants with perfectly inelastic demand to prices and income, and those zero consumption participants who report zero consumption at the time but who may consume once the price is right, for example. The former may relate to personal demographics and socioeconomic status, whilst the latter group may exhibit behaviour similar to other non-zero users and be more responsive to economic factors such as prices and income.

Let r denote a binary variable indicating the split between Regime 0 (with $r = 0$ for non-participants) and Regime 1 (with $r = 1$ for participants). r is related to a latent variable r^* via the mapping: $r = 1$ for $r^* > 0$ and $r = 0$ for $r^* \leqslant 0$. The latent variable r^* represents the propensity for participation and is given by

$$r^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $\mathbf{x}$ is a vector of covariates that determine the choice between the two regimes, $\boldsymbol{\beta}$ a vector of unknown coefficients, and ε a standard-normally distributed error term. Accordingly, the probability of an individual being in Regime 1 is given by (Maddala, 1983)

$$\Pr(r = 1|\mathbf{x}) = \Pr(r^* > 0|\mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta}), \tag{2}$$

where $\Phi(.)$ is the cumulative distribution function (*c.d.f.*) of the univariate standard normal distribution.

Conditional on $r = 1$, consumption levels under Regime 1 are represented by a discrete variable $\widetilde{y}$ ($\widetilde{y} = 0, 1, \ldots, J$) that is generated by an OP model via a second underlying latent variable $\widetilde{y}^*$:

$$\widetilde{y}^* = \mathbf{z}'\gamma + u, \tag{3}$$

with $\mathbf{z}$ being a vector of explanatory variables with unknown weights γ, and u an error term following a standard normal distribution. The mapping between $\widetilde{y}^*$ and $\widetilde{y}$ is given by

$$\widetilde{y} = \begin{cases} 0 & \text{if } \widetilde{y}^* \leqslant 0, \\ j & \text{if } \mu_{j-1} < \widetilde{y}^* \leqslant \mu_j \ \ (j = 1, \ldots, J-1), \\ J & \text{if } \mu_{J-1} \leqslant \widetilde{y}^*, \end{cases} \tag{4}$$

where $\mu_j (j = 1, \ldots, J-1)$ are boundary parameters to be estimated in addition to γ, and we assume throughout the paper that $\mu_0 = 0$. Note that, importantly, Regime 1 also allows for zero consumption. Also, there is no requirement that $\mathbf{x} = \mathbf{z}$. Under the assumption that

u is standard Gaussian, the OP probabilities are given by (Maddala, 1983)

$$\Pr(\tilde{y}) = \begin{cases} \Pr(\tilde{y} = 0|\mathbf{z}, r = 1) = \Phi(-\mathbf{z}'\gamma), \\ \Pr(\tilde{y} = j|\mathbf{z}, r = 1) = \Phi(\mu_j - \mathbf{z}'\gamma) - \Phi(\mu_{j-1} - \mathbf{z}'\gamma) \quad (j = 1, \ldots, J-1), \\ \Pr(\tilde{y} = J|\mathbf{z}, r = 1) = 1 - \Phi(\mu_{J-1} - \mathbf{z}'\gamma). \end{cases} \tag{5}$$

While r and $\tilde{y}$ are not individually observable in terms of the zeros, they are observed via the criterion

$$y = r\tilde{y}. \tag{6}$$

That is, to observe a $y = 0$ outcome we require either that $r = 0$ (the individual is a non-participant) or jointly that $r = 1$ and $\tilde{y} = 0$ (the individual is a zero consumption participant). To observe a positive y, we require jointly that the individual is a participant ($r = 1$) *and* that $\tilde{y}^* > 0$. Under the assumption that ε and u identically and independently follow standard Gaussian distributions, the full probabilities for y are given by

$$\Pr(y) = \begin{cases} \Pr(y = 0|\mathbf{z}, \mathbf{x}) = \Pr(r = 0|\mathbf{x}) + \Pr(r = 1|\mathbf{x}) \Pr(\tilde{y} = 0|\mathbf{z}, r = 1) \\ \Pr(y = j|\mathbf{z}, \mathbf{x}) = \Pr(r = 1|\mathbf{x}) \Pr(\tilde{y} = j|\mathbf{z}, r = 1) \quad (j = 1, \ldots, J) \end{cases}$$
$$= \begin{cases} \Pr(y = 0|\mathbf{z}, \mathbf{x}) = [1 - \Phi(\mathbf{x}'\boldsymbol{\beta})] + \Phi(\mathbf{x}'\boldsymbol{\beta})\Phi(-\mathbf{z}'\gamma) \\ \Pr(y = j|\mathbf{z}, \mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta})[\Phi(\mu_j - \mathbf{z}'\gamma) - \Phi(\mu_{j-1} - \mathbf{z}'\gamma)] \quad (j = 1, \ldots, J-1) \\ \Pr(y = J|\mathbf{z}, \mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta})[1 - \Phi(\mu_{J-1} - \mathbf{z}'\gamma)]. \end{cases} \tag{7}$$

In this way, the probability of a zero observation has been "inflated" as it is a combination of the probability of "zero consumption" from the OP process plus the probability of "non-participation" from the split probit model. Note that this specification is analogous to the zero-inflated/augmented count models, and that there may or may not be overlaps with the variables in **x** and **z**. Moreover, the model is also directly comparable to the double-hurdle limited dependent variable models (see, for example, Cragg, 1971).

Once the full set of probabilities has been specified and given an *i.i.d.* sample of size N from the population on $(y_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, N$, the parameters of the full model $\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma', \boldsymbol{\mu}')'$ can be consistently and efficiently estimated using maximum likelihood (ML) criteria, yielding asymptotically normally distributed maximum likelihood estimates (MLEs).[1] The log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=0}^{J} h_{ij} \ln[\Pr(y_i = j|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta})], \tag{8}$$

where the indicator function h_{ij} is

$$h_{ij} = \begin{cases} 1 & \text{if individual } i \text{ chooses outcome } j \\ 0 & \text{otherwise.} \end{cases} \quad (i = 1, \ldots, N;\ j = 0, 1, \ldots, J) \tag{9}$$

[1] An anonymous referee pointed out that an interesting line of future research would be to estimate such a model by Bayesian Monte Carlo Markov Chain techniques.

2.2. Generalising the model to correlated disturbances (ZIOPC)

As described above, the observed realisation of the random variable y can be viewed as the result of two separate latent equations, Eqs. (1) and (3), with uncorrelated error terms. However, these correspond to the same individual so it is likely that the two stochastic terms ε and u will be related. We now extend the model to have (ε, u) follow a bivariate normal distribution with correlation coefficient ρ, whilst maintaining the identifying assumption of unit variances. The full observability criteria are thus

$$y = r\widetilde{y} = \begin{cases} 0 & \text{if } (r^* \leqslant 0) \text{ or } (r^* > 0 \text{ and } \widetilde{y}^* \leqslant 0), \\ j & \text{if } (r^* > 0 \text{ and } \mu_{j-1} < \widetilde{y}^* \leqslant \mu_j) \quad (j = 1, \ldots, J-1), \\ J & \text{if } (r^* > 0 \text{ and } \mu_{J-1} < \widetilde{y}^*), \end{cases} \tag{10}$$

which translate into the following expressions for the probabilities:

$$\Pr(y) = \begin{cases} \Pr(y = 0|\mathbf{z}, \mathbf{x}) = [1 - \Phi(\mathbf{x}'\boldsymbol{\beta})] + \Phi_2(\mathbf{x}'\boldsymbol{\beta}, -\mathbf{z}'\boldsymbol{\gamma}; -\rho), \\ \Pr(y = j|\mathbf{z}, \mathbf{x}) = \Phi_2(\mathbf{x}'\boldsymbol{\beta}, \mu_j - \mathbf{z}'\boldsymbol{\gamma}; -\rho) - \Phi_2(\mathbf{x}'\boldsymbol{\beta}, \mu_{j-1} - \mathbf{z}'\boldsymbol{\gamma}; -\rho) \\ \quad (j = 1, \ldots, J-1), \\ \Pr(y = J|\mathbf{z}, \mathbf{x}) = \Phi_2(\mathbf{x}'\boldsymbol{\beta}, \mathbf{z}'\boldsymbol{\gamma} - \mu_{J-1}; \rho), \end{cases} \tag{11}$$

where $\Phi_2(a, b; \lambda)$ denotes the *c.d.f.* of the standardised bivariate normal distribution with correlation coefficient λ between the two univariate random elements.

ML estimation would again involve maxmisation of Eq. (8) replacing the probabilities of (7) with those of (11) and re-defining $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\mu}', \rho)'$. A *Wald* test of $\rho = 0$ is a test for independence of the two error terms and thus a test of the more general model given by Eq. (11) against the null of a simpler nested model of Eq. (7).[2]

2.3. Marginal effects

There are several sets of marginal effects that may be of interest in this model. For example, one may be interested in the marginal effects of an explanatory variable on the probability of "participation", $\Pr(r = 1)$, or the probabilities for the levels of consumption *conditional* on participation, $\Pr(\widetilde{y} = j|r = 1)$, or on the overall probabilities for different levels of consumption, $\Pr(y = j)$. In particular, the marginal effect on the overall probability of observing zero consumption, $\Pr(y = 0)$, is the sum of the effects on the probabilities of the two types of zeros; that is, the probability of non-participation and the probability of zero-consumption arising from participants who are infrequent or potential consumers.

The marginal effect of a dummy variable can be calculated as the difference in the probability of interest with the relevant dummy variable turned "on" and "off", conditional on given values of all other covariates. Note that the explanatory variable of interest may appear in only one of $\mathbf{x}$ or $\mathbf{z}$, or in both. For a continuous variable $\mathbf{x}_k$, the marginal effect on the participation probability in Eq. (2), which only relates to

[2] An OP model in Eqs. (3) and (4) with $\widetilde{y} \equiv y$ could be used as starting values for $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$. Those for $\boldsymbol{\beta}$ can be obtained from estimating a binary probit model defined by (1) and (2) with $\Pr(r = 1|\mathbf{x}) \equiv \Pr(y > 0|\mathbf{x})$. With fixed $\widehat{\boldsymbol{\theta}}_{\text{ZIOP}}$, a grid-search over $(-0.9, 0.9)$ would provide a start value for ρ.

explanatory variables in $\mathbf{x}$, is given by

$$\underset{\Pr(r=1)}{ME} = \frac{\partial \Pr(r=1)}{\partial x_k} = \phi(\mathbf{x}'\boldsymbol{\beta})\beta_k. \tag{12}$$

To derive the marginal effects on the overall probabilities for the general model of ZIOPC, we partition the explanatory variables and the associated coefficients as

$$\mathbf{x} = \begin{pmatrix} \mathbf{w} \\ \widetilde{\mathbf{x}} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{\mathbf{w}} \\ \widetilde{\boldsymbol{\beta}} \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} \mathbf{w} \\ \widetilde{\mathbf{z}} \end{pmatrix} \quad \text{and} \quad \gamma = \begin{pmatrix} \gamma_{\mathbf{w}} \\ \widetilde{\gamma} \end{pmatrix}, \tag{13}$$

where $\mathbf{w}$ represents the common variables that appear in both $\mathbf{x}$ and $\mathbf{z}$, with the associated coefficients $\boldsymbol{\beta}_{\mathbf{w}}$ and $\gamma_{\mathbf{w}}$ for the participation and consumption equations, respectively. $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{z}}$ denote those distinct variables that only appear in one of the latent equations, with $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\gamma}$ as their associated coefficients for the two equations.

Denote the unique explanatory variables for the whole model as $\mathbf{x}^* = (\mathbf{w}', \widetilde{\mathbf{x}}', \widetilde{\mathbf{z}}')'$, and set the associated coefficient vectors for $\mathbf{x}^*$ as $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_w', \widetilde{\boldsymbol{\beta}}', 0')'$ and $\gamma^* = (\gamma_w', 0', \widetilde{\gamma}')'$. The marginal effects of the explanatory variable vector $\mathbf{x}^*$ on the full probabilities in Eq. (11) are given by

$$\begin{aligned}
\underset{\Pr(y=0)}{\mathbf{ME}} &= \left[\Phi\left(\frac{-\mathbf{z}'\gamma + \rho\mathbf{x}'\boldsymbol{\beta}}{\sqrt{1-\rho^2}}\right) - 1\right]\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}^* - \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} - \rho\mathbf{z}'\gamma}{\sqrt{1-\rho^2}}\right)\phi(\mathbf{z}'\gamma)\gamma^*, \\
\underset{\Pr(y=0)}{\mathbf{ME}} &= \left[\Phi\left(\frac{-\mathbf{z}'\gamma + \rho\mathbf{x}'\boldsymbol{\beta}}{\sqrt{1-\rho^2}}\right) - 1\right]\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}^* - \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} - \rho\mathbf{z}'\gamma}{\sqrt{1-\rho^2}}\right)\phi(\mathbf{z}'\gamma)\gamma^* \\
&\quad + \left[\phi(\mathbf{z}'\gamma)\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} - \rho\mathbf{z}'\gamma}{\sqrt{1-\rho^2}}\right) - \phi(\mu_1 - \mathbf{z}'\gamma)\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} + \rho(\mu_1 - \mathbf{z}'\gamma)}{\sqrt{1-\rho^2}}\right)\right]\gamma^*, \\
\underset{\Pr(y=2)}{\mathbf{ME}} &= \left[\Phi\left(\frac{\mu_2 - \mathbf{z}'\gamma + \rho\mathbf{x}'\boldsymbol{\beta}}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{\mu_1 - \mathbf{z}'\gamma + \rho\mathbf{x}'\boldsymbol{\beta}}{\sqrt{1-\rho^2}}\right)\right]\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}^* \\
&\quad + \left[\phi(\mu_1 - \mathbf{z}'\gamma)\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} + \rho(\mu_1 - \mathbf{z}'\gamma)}{\sqrt{1-\rho^2}}\right) - \phi(\mu_2 - \mathbf{z}'\gamma)\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} + \rho(\mu_2 - \mathbf{z}'\gamma)}{\sqrt{1-\rho^2}}\right)\right]\gamma^*, \\
&\ \vdots \\
\underset{\Pr(y=J)}{\mathbf{ME}} &= \Phi\left(\frac{\mathbf{z}'\gamma - \mu_{J-1} - \rho\mathbf{x}'\boldsymbol{\beta}}{\sqrt{1-\rho^2}}\right)\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}^* + \phi(\mathbf{z}'\gamma - \mu_{J-1}) \\
&\quad \times\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta} - \rho(\mathbf{z}'\gamma - \mu_{J-1})}{\sqrt{1-\rho^2}}\right)\gamma^*,
\end{aligned} \tag{14}$$

where $\phi(.)$ is the *p.d.f.* of the standard univariate normal distribution. Note that the marginal effect on $\Pr(y=0)$ can be decomposed into the marginal effects on the probabilities of the two types of zeros. Marginal effects for the ZIOP model are obtained as above but with $\rho = 0$. Standard errors of the marginal effects can be obtained by the Delta method (see, for example, Greene, 2003, pp. 674–675). An alternative method is to use simulated asymptotic sampling techniques. Specifically, randomly draw $\boldsymbol{\theta}$ from $MVN(\widehat{\boldsymbol{\theta}}, Var[\widehat{\boldsymbol{\theta}}])$ a large number of times. For each random draw, calculate the marginal effects using either the analytical expressions of Eq. (14) or the numerical derivatives of the probability expressions. The empirical standard deviations of the simulated marginal effects are valid asymptotic estimates of their standard errors.[3]

2.4. Hypothesis testing and model selection issues

Testing between the ZIOP and ZIOPC models can be based on a simple t-test of $\rho = 0$, using the standard errors from the estimated Hessian. With regard to the ZIOP (or ZIOPC) *versus* the OP model, they are not nested in the usual sense of parameter restrictions. The ZIOP (or ZIOPC) model becomes an OP model when $\Pr(r = 1|\mathbf{x}) \equiv 1$, or $\mathbf{x}'\beta \to \infty$ in Eq. (2), implying all individuals are in Regime 1 and there is no "zero-splitting" process. Although having two non-nested models in this case, a generalised *likelihood ratio* (*LR*) statistic could be used, with degrees of freedom being given by the number of additional parameters estimated in the more general model. The *LR* test is known to have good properties in non-standard testing problems (see, for example, Andrews and Ploberger, 1995; Chesher and Smith, 1997). Similarly lacking theoretical underpinnings in this situation but a useful general specification test in many situations, the *Hausman* specification test (Hausman, 1978) could also be considered, with the degrees of freedom being the number of common parameters estimated in the competing models. As indicated in the Monte Carlo results in Section 3, both tests actually perform quite well.

A more theoretically based approach is the *Vuong* test (Vuong, 1989) for testing between two non-nested models, which has been suggested in the related context of testing a zero-inflated Poisson *versus* a simple count model (Greene, 2003). Denote $f_h(y_i|\mathbf{x}_i, z_i)$ as the predicted probability using Model h ($h = 1$ and 2 for OP and ZIOP/ZIOPC) that y_i equals the random variable y equals the observed y_i and let

$$m_i = \log\left(\frac{f_1(y_i|\mathbf{x}_i, z_i)}{f_2(y_i|\mathbf{x}_i, z_i)}\right). \tag{15}$$

To test the null hypothesis that $E(m_i) = 0$, or that there is no difference in the probabilities of correct prediction using the two models, the *Vuong* statistic is given by

$$\upsilon = \frac{\sqrt{N}(1/N\sum_{i=1}^{N} m_i)}{\sqrt{1/N\sum_{i=1}^{N}(m_i - \bar{m})^2}}, \tag{16}$$

which has a standard normal limiting distribution. The test statistic is bidirectional in the sense that $|\upsilon| < 1.96$ lends support to neither model, whereas $\upsilon < -1.96$ favours Model 2 and $\upsilon > 1.96$ favours Model 1 (Vuong, 1989).

Finally, in such a non-nested situation, information based model-selection criteria, such as *AIC*, *BIC* and consistent AIC (*CAIC*), are appropriate for choosing between alternative models. These are given by $AIC = -2\ell(\boldsymbol{\theta}) + k$, $BIC = -2\ell(\boldsymbol{\theta}) + (\ln N)k$, and $CAIC = -2\ell(\boldsymbol{\theta}) + (1 + \ln N)k$ (see, for example, Cameron and Trivedi, 1998, p. 183), where k is the total number of parameters estimated and $\ell(\boldsymbol{\theta})$ the maximised log-likelihood function. The preferred model is that with the smallest value.

[3] The Delta and the simulation methods represent two distinct asymptotic approximations. Nearly identical results were obtained from the two approaches, indicating the adequacy of the underlying asymptotic theory in this case.

3. Finite sample performance

Although the model satisfies the regularity conditions for ML estimation (see Greene, 2003), results of some Monte Carlo experiments are presented in this section to provide some evidence on the finite sample performance of the ML estimator and the model selection criteria. We consider, in turn, the cases when the data generating process (*d.g.p.*) is ZIOPC, ZIOP and OP. Estimation was undertaken in the *Gauss* matrix programming language, using the CML maximum likelihood estimation add-in module.[4]

3.1. Performance under ZIOPC/ZIOP

3.1.1. Monte Carlo design

$R = 1{,}000$ repeated samples, each with a sample size of $N = 1{,}000$, are generated from a ZIOPC *d.g.p.*, and all three models of ZIOPC, ZIOP and OP are estimated (Experiment 1). This is then repeated for a *d.g.p.* of ZIOP with $\rho = 0$ (Experiment 2). We draw $\mathbf{x}$ from $\mathbf{x} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2)'$, where $\mathbf{x}_1 = \log(\mathit{Uniform}[0, 100])$ and $\mathbf{x}_2 = \mathbf{1}_{\{\mathit{Uniform}[0,1]>0.25\}}$. Observations for $\mathbf{z}$ are generated from $\mathbf{z} = (\mathbf{1}, \mathbf{z}_1)'$, where $\mathbf{z}_1 \equiv \mathbf{x}_1$. We have chosen the continuous variable $\mathbf{z}_1 \equiv \mathbf{x}_1$ to mimic variables such as age and income, and the dummy variable, $\mathbf{x}_2$ to represent qualitative characteristics such as gender or marital status. The set of $N = 1{,}000$ draws of $\mathbf{x}$ and $\mathbf{z}$ is generated once and subsequently held fixed.

Parameter values for both experiments are set as follows: $(\beta_0, \beta_1, \beta_2)' = (1, -0.25, -1)'$, $(\gamma_0, \gamma_1)' = (0.5, 1)'$, $(\mu_1, \mu_2)' = (4.5, 5.5)'$, and for Experiment 1, $\rho = 0.5$. Note that we have set the coefficients β_1 and γ_1 as having opposite signs allowing for the same explanatory variable ($\mathbf{x}_1 \equiv \mathbf{z}_1$) to have opposing effects on the two latent variables. The parameter values are also chosen to yield around 70% of zero observations.

3.1.2. Monte Carlo results

The results are summarised in Tables 1 and 2. As individual coefficients in such discrete models do not convey much information, we present the results in terms of estimated marginal effects evaluated at sample means of the observed covariates. For each of the $j = 0, \ldots, J$ outcomes, we present the true marginal effects, the estimated marginal effects averaged over the R runs ($\overline{ME}$), the root mean square error of the R estimated marginal effects relative to the true *ME*s (*RMSE*), as well as the empirical coverage probabilities (*CP*), measured as the percentages of times the true marginal effects fall within the estimated 95% confidence intervals. We also present the same set of information for the estimated parameters $\boldsymbol{\mu}$ and ρ, though we do not present results for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.[5]

Some model selection and summary statistics are also presented. $\overline{RMSE\_P}$ refers to the root mean square error of the predicted probabilities for all outcomes and observations, averaged over the R Monte Carlo runs, where for the mth replication $RMSE_m = \sqrt{(1/NJ)\sum_{i=1}^{N}\sum_{j=0}^{J}(\widehat{P}_{ij}^{m} - P_{ij}^{m})^2}$, $(m = 1, \ldots, R)$. *Correct* gives the average percentage of correct predictions for y based on the maximum probability rule, whilst *Time* refers to the average estimation time (in minutes). The results for the *Wald* test of the

[4] Code is available from the authors on request. Also the model will be available in the next release of *Limdep* (version 9.0), *NLOGIT* (4.0).

[5] Results for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be found in Harris and Zhao (2004).

Table 1
Monte Carlo results under ZIOPC: Experiment 1

			$\Pr(y=0\|\bar{\mathbf{x}},\bar{\mathbf{z}})$				$\Pr(y=1\|\bar{\mathbf{x}},\bar{\mathbf{z}})$		
		TRUE	OP	ZIOP	ZIOPC	TRUE	OP	ZIOP	ZIOPC
Marginal effects (*ME*)									
$x_1 = z_1$	$\overline{ME}$	0.080	0.013	0.083	0.082	−0.149	−0.004	−0.147	−0.150
	RMSE		(0.068)	(0.018)	(0.018)		(0.145)	(0.015)	(0.016)
	CP		0.003	0.955	0.955		0.000	1.000	1.000
x_2	$\overline{ME}$	0.320		0.321	0.322	−0.165		−0.134	−0.165
	RMSE			(0.031)	(0.031)			(0.035)	(0.022)
	CP			0.957	0.956			1.000	1.000
			$\Pr(y=2\|\bar{\mathbf{x}},\bar{\mathbf{z}})$				$\Pr(y=3\|\bar{\mathbf{x}},\bar{\mathbf{z}})$		
		TRUE	OP	ZIOP	ZIOPC	TRUE	OP	ZIOP	ZIOPC
$x_1 = z_1$	$\overline{ME}$	−0.001	−0.004	0.002	−0.002	0.071	−0.005	0.063	0.070
	RMSE		(0.007)	(0.011)	(0.012)		(0.076)	(0.012)	(0.010)
	CP		0.997	1.000	1.000		0.000	1.000	1.000
x_2	$\overline{ME}$	−0.118		−0.126	−0.119	−0.037		−0.062	−0.038
	RMSE			(0.019)	(0.019)			(0.027)	(0.013)
	CP			0.998	1.000			1.000	1.000
		TRUE	OP	ZIOP	ZIOPC				
Coefficients									
μ_1	$\overline{\mu_1}$	4.500	0.446	4.787	4.461				
	RMSE		(4.050)	(0.860)	(0.830)				
	CP		0.000	0.976	0.947				
μ_2	$\overline{\mu_2}$	5.500	0.893	5.872	5.460				
	RMSE		(4.610)	(0.910)	(0.860)				
	CP		0.000	0.976	0.947				
ρ	$\overline{\rho}$	0.500			0.482				
	RMSE				(0.170)				
	CP				0.927				
		OP	ZIOP	ZIOPC					
$\overline{RMSE\_P}$		0.116	0.023	0.020					
		(0.001)	(0.005)	(0.006)					
Correct		0.731	0.744	0.744					
Time		0.025	0.075	0.493					
Wald			0.235	0.765					
Vuong		0.00	1.00						
Vuong(*C*)		0.00		1.00					
LR			1.00						
LR(*C*)				1.00					
Hausman			1.00						
Hausman(*C*)				0.99					
AIC		0.000	0.074	0.926					
BIC		0.000	0.540	0.460					
CAIC		0.000	0.589	0.411					

Table 2
Monte Carlo results under ZIOP: Experiment 2

			$\Pr(y=0 \mid \mathbf{z},\mathbf{x})$				$\Pr(y=1 \mid \mathbf{z},\mathbf{x})$		
		TRUE	OP	ZIOP	ZIOPC	TRUE	OP	ZIOP	ZIOPC
Marginal effects									
$x_1 = z_1$	$\overline{ME}$	0.080	0.018	0.083	0.083	−0.146	−0.009	−0.146	−0.147
	RMSE		(0.063)	(0.019)	(0.019)		(0.137)	(0.017)	(0.017)
	CP		0.003	0.953	0.955		0.000	0.979	0.982
x_2	$\overline{ME}$	0.320		0.320	0.320	−0.206		−0.206	−0.206
	RMSE			(0.033)	(0.032)			(0.024)	(0.027)
	CP			0.951	0.952			1.000	1.000
			$\Pr(y=2 \mid \mathbf{z},\mathbf{x})$				$\Pr(y=3 \mid \mathbf{z},\mathbf{x})$		
		TRUE	OP	ZIOP	ZIOPC	TRUE	OP	ZIOP	ZIOPC
$x_1 = z_1$	$\overline{ME}$	0.033	−0.005	0.032	0.032	0.033	−0.004	0.032	0.032
	RMSE		(0.038)	(0.010)	(0.010)		(0.037)	(0.006)	(0.006)
	CP		0.000	0.996	0.997		0.000	1.000	1.000
x_2	$\overline{ME}$	−0.087		−0.087	−0.087	−0.028		−0.028	−0.027
	RMSE			(0.013)	(0.015)			(0.007)	(0.008)
	CP			1.000	1.000			1.000	1.000
		TRUE	OP	ZIOP	ZIOPC				
Coefficients									
μ_1	$\overline{\mu_1}$	4.500	0.667	4.449	4.393				
	RMSE		(3.830)	(0.780)	(0.780)				
	CP		0.000	0.943	0.939				
μ_2	$\overline{\mu_2}$	5.500	1.206	5.454	5.383				
	RMSE		(4.290)	(0.790)	(0.800)				
	CP		0.000	0.943	0.939				
ρ	$\overline{\rho}$	0.000			0.005				
	RMSE				(0.220)				
	CP				0.927				
		OP	ZIOP	ZIOPC					
$\overline{RMSE\_P}$		0.110	0.019	0.019					
		(0.002)	(0.006)	(0.006)					
Correct		0.733	0.748	0.748					
Time		0.0262	0.0811	0.3184					
Wald			0.927	0.073					
Vuong		0.00	1.00						
Vuong(*C*)		0.00		1.00					
LR			1.00						
LR(*C*)				1.00					
Hausman			1.00						
Hausman(*C*)				1.00					
AIC		0.000	0.670	0.330					
BIC		0.000	0.988	0.012					
CAIC		0.000	0.991	0.009					

null of ZIOP (H_0: $\rho = 0$) against the alternative of ZIOPC are given as the percentage of times that the statistic lends support to each model. *LR*/*LR*(*C*) and *Hausman*/*Hausman*(*C*) are, respectively, the percentage of times the *LR* and *Hausman* statistics favour the ZIOP (ZIOPC) over the OP model. *Vuong*/*Vuong*(*C*) corresponds to Vuong's (1989) non-nested test as applied to the ZIOP/ZIOPC model *versus* OP, again expressed as the percentage of times each model is selected. Finally, the percentage of times each of the *AIC*, *BIC* and *CAIC* selects each of the three models is also reported. All tests are undertaken at 5% nominal size.

As can be seen from Table 1, when the true model is ZIOPC (Experiment 1) and a simple OP model is estimated, not surprisingly, both the estimated marginal effects and the estimated boundary parameters are severely biased. With the exception of one outcome, 95% empirical *CP* are essentially zero. On the other hand, estimation of a ZIOP model ignoring the correlation performs quite well. Average estimated marginal effects are very close to the true ones, and moreover have small *RMSE*s. Allowing for the (true) correlation in estimation (ZIOPC) further improves the results. *RMSE* measures for ZIOPC are even better with essentially zero biases and coverage probabilities ranging from 0.927 to 1.[6] The average estimate of ρ is 0.48 compared to the actual value of 0.5, with an empirical *CP* of 93%.

In terms of correctly estimating probabilities, the OP clearly fares poorly with an average *RMSE* of 0.116. Significant improvements are afforded by the ZIOP and ZIOPC models; here the mean *RMSE* falls sharply to 0.023 and 0.020, respectively. The percentage of correct predictions is fairly similar across all models. This is the result of a common phenomenon typical in discrete choice models, where models tend to simply predominantly predict the most frequently observed outcome (here zeros), resulting in poor "predictive" performance.

For applied researchers, an important issue is a model selection procedure to correctly choose between alternative models. As shown in Table 1, the *Wald* test of ZIOPC *versus* ZIOP correctly rejects the null and selects the correct ZIOPC model in 77% of cases. For choosing between the non-nested models, Vuong's (1989) statistic correctly selects the ZIOPC model over the incorrect OP in all cases. The uncorrelated ZIOP model is also preferred to the OP model by the *Vuong* test in all instances. The *LR* statistics also correctly reject the OP model all of the time. The *Hausman* statistics fare similarly well, only incorrectly rejecting in 1% of instances (for the correlated version). In terms of the information criteria, in no instances do any of the criteria incorrectly select the OP model. *AIC* significantly favours the ZIOPC model, whereas *BIC* and *CIAC* have an approximate equal split in choosing between the ZIOP and ZIOPC models. However, as already stated, a preferable method of choosing between these two nested models would be a *Wald* test of $\rho = 0$. Finally, with regard to estimation times, the ZIOPC (with an average of 0.493 min per estimation) appears to be more computationally intensive than the simpler OP (0.025 min) and its uncorrelated counterpart ZIOP (0.075 min).

Turning to Experiment 2 in Table 2 where the *d.g.p.* is ZIOP (with $\rho = 0$), here we would expect the OP to fare poorly and the ZIOP and ZIOPC to excel, with the estimates of ρ in the latter to be "small" and insignificant. Indeed, the OP estimated marginal effects and boundary parameters are again quite severely biased, with *CP* of 0.003 or lower, whereas

[6] Note that all of these *CP* are based on asymptotic distributions whilst *N* here is "small" at 1,000. Empirical *CP* would therefore be much closer to the theoretical 0.95 values for larger sample sizes.

both of the ZIOP and ZIOPC ones are essentially identical with small biases and empirical *CP* ranging from 0.939 to 1. The average estimate of ρ is 0.005 and at 5% nominal size one would incorrectly reject the null of $\rho = 0$ in 7.3% of cases. The ZIOP and ZIOPC models clearly dominate the misspecified simple OP model in terms of $\overline{RMSE\_P}$. Once more, in all instances the *Vuong* statistic correctly chooses against the simple OP model, as do the *LR* and *Hausman* statistics. With regard to model selection criteria, the OP is never chosen by any of the criteria, and here the information criteria perform much better with regard to choosing the correct model.

3.1.3. Exclusion restrictions

It is often the case in such two-part models that precision of parameter estimates is enhanced if there are explicit exclusion restrictions in the specification of the covariates in the two equations. For example, in the well-known Heckman-selection equation (Heckman, 1979), although the correlation between the selection and regression equations is identified by the nonlinearities involved, due to multicollinearity concerns, this correlation is often imprecisely estimated if $\mathbf{x} \equiv \mathbf{z}$. Smith (2003) also suggests that identification of the correlation parameter may be stronger in samples with more than 50% zeros. To examine the likely effect of exclusion restrictions, Experiment 1 with a *d.g.p.* of ZIOPC is re-run assuming $\mathbf{x} \equiv \mathbf{z}$. The results are presented in Table 3.

Here all results for the marginal effects are somewhat similar to Experiment 1 where exclusion restrictions were in place. All of the model selection procedures also invariably correctly select the larger models over OP. However, there is indeed evidence that the correlation coefficient, ρ, is not properly identified; the average estimate for ρ is only 0.038 compared to the true value of 0.5. Based on the *Wald* statistic, in only 2.3% of the cases would one correctly select the ZIOPC model over the ZIOP variant. Furthermore, convergence problems were encountered for particular draws of the random variables within the Monte Carlo experiment. Indeed, Smith (2003) has also suggested that weak identification can lead to computational problems such as lack of convergence in similar models.

From an empirical point of view however, given that the zeros are assumed to come from two different regimes, a model with $\mathbf{x} \equiv \mathbf{z}$ is not going to be a scenario that an applied researcher would necessarily entertain.

3.2. Performance under OP

We now consider the case when the true model is, in fact, the usual OP model. In this case, even though $\mathbf{x}$ and $\boldsymbol{\beta}$ do not feature in the true OP *d.g.p.*, ZIOP and ZIOPC models were estimated *as if* they did. We consider several scenarios for the explanatory variables $\mathbf{x}$ and $\mathbf{z}$. Experiment 4 has partly overlapping $\mathbf{x}$ and $\mathbf{z}$ as in Experiments 1 and 2, with $\mathbf{x} = \{1, \log(\mathit{Uniform}[0, 100]), \mathbf{1}_{\{\text{Uniform}[0,1] > 0.25\}}\}$ and $\mathbf{z}$ equal to the first two columns of $\mathbf{x}$. Experiment 5 assumes that $\mathbf{x}$ and $\mathbf{z}$ have *explicit* exclusion restrictions and no overlapping variables, with $\mathbf{z} = \{1, |\mathrm{N}(0, 4)|\}$ and $\mathbf{x}$ as before. In Experiment 6, we consider a case of complete overlap, with $\mathbf{x} \equiv \mathbf{z} = \{1, \log(\mathit{Uniform}[0, 100])\}$. Finally Experiment 7 has $\mathbf{x}$ and $\mathbf{z}$ as in Experiment 4 except that $\mathbf{x}$ has an additional $\mathrm{N}(0, 4)$ variate. Again $R = 1{,}000$ Monte Carlo replications are considered in all scenarios.

The marginal effect results from these experiments are in Table 4, and summary statistics and model selection results in Table 5. Convergence problems were encountered with the

Table 3
Monte Carlo results under ZIOPC ($x = z$): Experiment 3

			$\Pr(y=0\|\mathbf{z},\mathbf{x})$				$\Pr(y=1\|\mathbf{z},\mathbf{x})$		
		TRUE	OP	ZIOP	ZIOPC	TRUE	OP	ZIOP	ZIOPC
Marginal effects									
$x_1 = z_1$	$\overline{ME}$	0.099	−0.032	0.102	0.102	−0.301	0.005	−0.295	−0.297
	RMSE		(0.132)	(0.021)	(0.021)		(0.306)	(0.022)	(0.021)
	CP		0.000	0.952	0.959		0.000	0.986	0.999
			$\Pr(y=2\|\mathbf{z},\mathbf{x})$				$\Pr(y=3\|\mathbf{z},\mathbf{x})$		
		TRUE	OP	ZIOP	ZIOPC	TRUE	OP	ZIOP	ZIOPC
$x_1 = z_1$	$\overline{ME}$	0.079	0.012	0.071	0.073	0.123	0.015	0.123	0.123
	RMSE		(0.067)	(0.019)	(0.018)		(0.108)	(0.012)	(0.012)
	CP		0.000	1.000	0.997		0.000	1.000	1.000
		TRUE	OP	ZIOP	ZIOPC				
Coefficients									
μ_1	$\overline{\mu_1}$	4.500	0.694	4.865	4.826				
	RMSE		(3.810)	(0.720)	(0.730)				
	CP		0.000	0.949	0.986				
μ_2	$\overline{\mu_2}$	5.500	1.315	5.955	5.902				
	RMSE		(4.190)	(0.780)	(0.790)				
	CP		0.000	0.949	0.986				
ρ	$\overline{\rho}$	0.500			0.038				
	RMSE				(0.510)				
	CP				0.990				
			OP	ZIOP	ZIOPC				
$\overline{RMSE\_P}$			0.122	0.019	0.019				
			(0.002)	(0.006)	(0.006)				
Correct			0.470	0.523	0.523				
Time			0.027	0.067	0.344				
Wald				0.977	0.023				
Vuong			0.00	1.00					
Vuong(C)			0.00		1.00				
LR				1.00					
LR(C)					1.00				
Hausman				1.00					
Hausman(C)					0.94				
AIC			0.000	0.976	0.024				
BIC			0.000	0.997	0.003				
CAIC			0.000	0.998	0.002				

ZIOPC model in these experiments. For this reason, only the ZIOP model was estimated. For the applied researchers, if convergence problems are encountered with the ZIOPC, it may suggest that the data is inconsistent with a zero-splitting process.

Results in Table 4 show that when the true *d.g.p.* is OP, a ZIOP model actually performs very well. In fact, the average estimated marginal effects from both OP and ZIOP models

Table 4
Monte Carlo results under OP: Experiments 4–7—marginal effects

Experiment 4		TRUE	OP	ZIOP	TRUE	OP	ZIOP
		$\Pr(y=0\|\mathbf{z},\mathbf{x})$			$\Pr(y=1\|\mathbf{z},\mathbf{x})$		
$x_1=z_1$	$\overline{ME}$	0.000	0.000	0.000	−0.373	−0.376	−0.376
	CP		0.851	0.856		1.000	1.000
x_2	$\overline{ME}$	0.000		0.000	0.000		0.000
	CP			0.999			0.999
		$\Pr(y=2\|\mathbf{z},\mathbf{x})$			$\Pr(y=3\|\mathbf{z},\mathbf{x})$		
$x_1=z_1$	$\overline{ME}$	0.216	0.219	0.219	0.157	0.157	0.157
	CP		0.893	0.994		1.000	1.000
x_2	$\overline{ME}$	0.000		0.000	0.000		0.000
	CP			0.999			0.999
Experiment 5							
		$\Pr(y=0\|\mathbf{z},\mathbf{x})$			$\Pr(y=1\|\mathbf{z},\mathbf{x})$		
x_1	$\overline{ME}$	0.000		0.0001	0.000		−0.0001
	CP			0.999			0.999
z_1	$\overline{ME}$	−0.0412	−0.0412	−0.0388	0.0171	0.0171	0.0157
	CP		0.936	0.926		1	1
		$\Pr(y=2\|\mathbf{z},\mathbf{x})$			$\Pr(y=3\|\mathbf{z},\mathbf{x})$		
x_1	$\overline{ME}$	0.000		0	0.000		0
	CP			1			1
z_1	$\overline{ME}$	0.0227	0.0227	0.0218	0.0014	0.0015	0.0013
	CP		0.94	0.936		0.817	0.804
Experiment 6							
		$\Pr(y=0\|\mathbf{z},\mathbf{x})$			$\Pr(y=1\|\mathbf{z},\mathbf{x})$		
$x=z$	$\overline{ME}$	0.000	0.000	0.000	−0.373	−0.375	−0.375
	CP		0.859	0.906		0.999	1.000
		$\Pr(y=2\|\mathbf{z},\mathbf{x})$			$\Pr(y=3\|\mathbf{z},\mathbf{x})$		
$x=z$	$\overline{ME}$	0.216	0.217	0.218	0.157	0.158	0.157
	CP		0.924	0.999		1.000	1.000
Experiment 7							
		$\Pr(y=0\|\mathbf{z},\mathbf{x})$			$\Pr(y=1\|\mathbf{z},\mathbf{x})$		
$x_1=z_1$	$\overline{ME}$	0.000	0.000	0.000	−0.373	−0.374	−0.374
	CP		0.869	0.771		1.000	1.000
x_2	$\overline{ME}$	0.000		0.000	0.000		0.000
	CP			0.996			0.996
x_3	$\overline{ME}$	0.000		0.000	0.000		0.000
	CP			0.996			0.996
		$\Pr(y=2\|\mathbf{z},\mathbf{x})$			$\Pr(y=3\|\mathbf{z},\mathbf{x})$		
$x_1=z_1$	$\overline{ME}$	0.216	0.218	0.218	0.157	0.156	0.156
	CP		0.917	0.997		1.000	1.000
x_2	$\overline{ME}$	0.000		0.000	0.000		0.000
	CP			0.996			0.996
x_3	$\overline{ME}$	0.000		0.000	0.000		0.000
	CP			0.996			0.996

Table 5
Monte Carlo results under OP: Experiments 4–7—summary statistics[a]

	Experiment 4		*Experiment* 5		*Experiment* 6		*Experiment* 7	
	OP	ZIOP	OP	ZIOP	OP	ZIOP	OP	ZIOP
$\overline{RMSE\_P}$	0.015	0.018	0.014	0.015	0.014	0.016	0.014	0.019
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
Correct	0.594	0.594	0.861	0.861	0.592	0.593	0.593	0.593
Vuong	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.03
LR		0.02		0.04		0.01		0.03
Hausman		0.01		0.01		0.01		0.01
AIC	0.791	0.209	0.743	0.257	0.834	0.166	0.724	0.276
BIC	1.000	0.000	0.998	0.002	1.000	0.000	1.000	0.000
CAIC	1.000	0.000	0.999	0.001	1.000	0.000	1.000	0.000
$\Pr(r=1\|\bar{\mathbf{x}})$	–	1.000	–	0.999	–	1.000	–	1.000
$\overline{\Pr(r=1\|\mathbf{x}_i)}$	–	0.993	–	0.998	–	0.994	–	0.993
95% range	–	(0.965,1)	–	(0.993,1)	–	(0.968,1)	–	(0.965,1)

[a] 95% Range refers to the empirical 95% range of $\Pr(r=1|\mathbf{x}_i)$.

are almost identical, both being very close to the true ones. Table 5 shows why this is the case; when the true data are OP, the zero split of the ZIOP in Eq. (2) rules that almost all observations are from Regime 1 and $\Pr(r=1|\mathbf{x}) \to 1$ ($\mathbf{x}'\boldsymbol{\beta} \to \infty$, $\Phi(\mathbf{x}'\boldsymbol{\beta}) \to 1$). As shown in Table 5, the average probability of $r=1$ evaluated at sample means, $\Pr(r=1|\bar{\mathbf{x}})$, is between 0.999 and 1 for the four experiments. This probability averaged across all individuals and all replications, $\overline{\Pr(r=1|x_i)}$, is also very close to unity, and the empirical 95% range of this latter probability is between 0.965 and 1 for all four experiments. This is a very favorable result indicating that even the (misspecified) ZIOP model is estimated when the true model is OP, the parameter estimates of $\boldsymbol{\beta}$ are such that the ZIOP reduces to OP even the two models are not nested in the usual sense.[7]

In terms of *RMSE* for the predicted probabilities, OP does slightly better than ZIOP. Again both models have near identical performance in the percentage of correct predictions. The *LR* statistic once more, somewhat surprisingly given its lack of theoretical justification, appears to work very well, with empirical sizes ranging from 1% to 4% (at nominal 5% level). Similarly the *Hausman* statistic is only marginally undersized with empirical sizes of 1%. The information criteria *BIC* and *CAIC* correctly choose the OP model in 99.8% or more of cases, while *AIC* does so in more than 72.4% of instances. In other words, the information criteria and *LR* and *Hausman* statistics appear to be able to choose the correct model when the true data are not from a zero split *d.g.p.*.

On the other hand, the *Vuong* statistic appears to be unable to choose between the two non-nested models. Recall that the test statistic is bidirectional. For the bulk of all experiments the test statistic falls in the "indeterminate" region ($|\upsilon| < 1.96$), leading to the conclusion that neither of the two models is preferred. For all experiments, there is not one single case where the true OP is chosen ($\upsilon > 1.96$), and for around 1–3% of the time the ZIOP is actually preferred. These results are confirmed in the Q–Q plots in Fig. 1, which

[7] This result was correctly conjectured by two anonymous referees, one of whom also suggested that as $N \to \infty$ so will $\widehat{\beta}_1$.

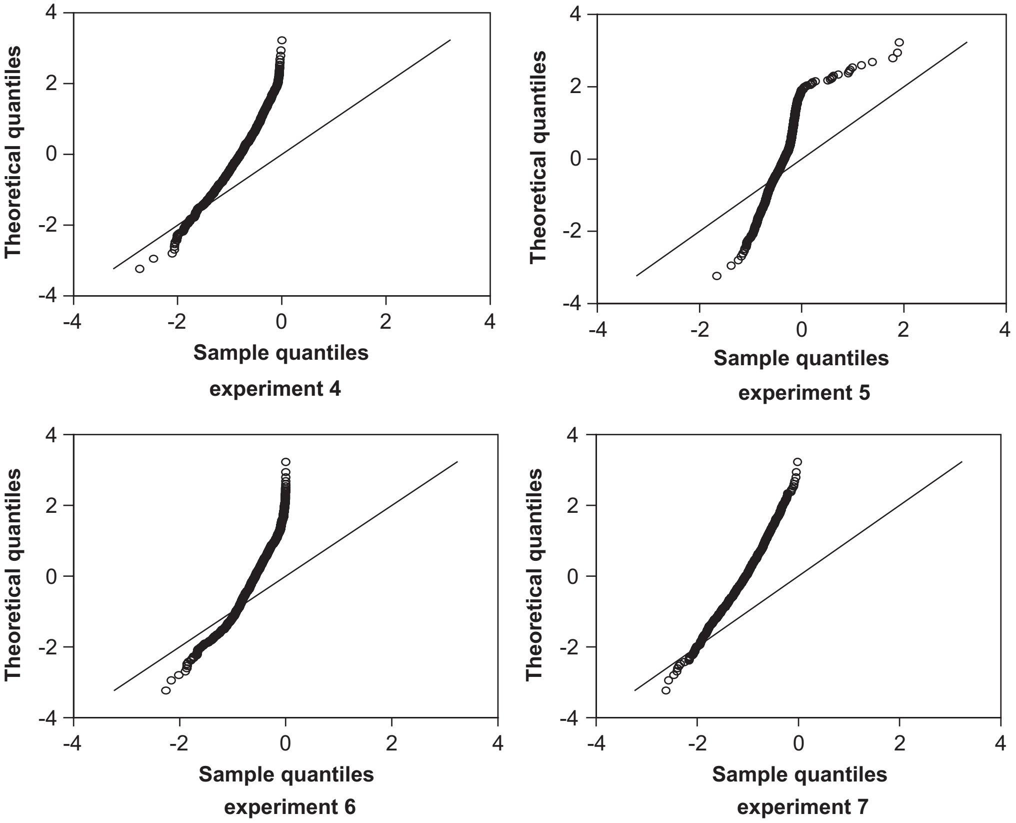

Fig. 1. Vuong statistic: theoretical versus empirical quantiles under OP.

contains plots of empirical quantiles of the *Vuong* statistic against theoretical ones (of a standard normal distribution) for the four experiments. For the sample size and parameter settings in these experiments, the plots are clearly not close to the 45° lines where the empirical and theoretical quantiles concur. In fact, for three of the four experiments, OP is unlikely to be chosen as the *Vuong* statistic seems to stay negative. The empirical 5% critical values for claiming the ZIOP model for Experiments 4–7 are -1.65, -0.86, -1.29 and -1.84, compared to the theoretical critical value of -1.96. We also note here that the empirical *Vuong* statistic in Experiments 1–3, where ZIOP models are the true *d.g.p.*, have large negative values that are below -5. This is shown in the Q–Q plots for Experiments 1–3 in Fig. 2.

In summary of the Monte Carlo results when the true *d.g.p.* is ZIOPC/ZIOP, both zero-inflated models perform well. The *Vuong* test, *LR* and *Hausman* statistics, as well as the information criteria should all correctly select the ZIOPC/ZIOP models. In choosing between the ZIOP and ZIOPC models, a standard t-test on the estimated value of ρ should be used. When the data has been generated according to an OP process, estimation of a ZIOP model will still yield accurate estimates of the quantities of interest, with the probability for Regime 1 tending to unity in the split decision such that ZIOP tends to OP. Moreover, the information criteria are likely to correctly select the smaller model (with

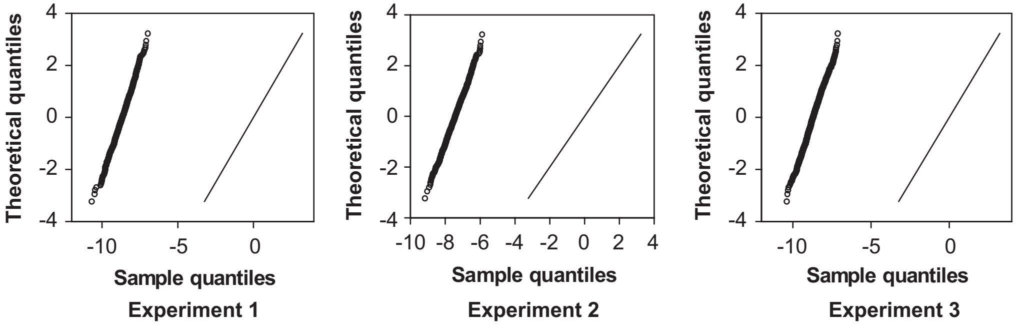

Fig. 2. Vuong statistic: theoretical versus empirical quantiles under ZIOP(C).

BIC and *CAIC* all of the time and *AIC* over 70% of the time). Similarly, somewhat surprisingly, both of the *Hausman* and *LR* tests have empirical sizes very close to nominal ones, with the latter having marginally better performance. There is evidence, however, that the *Vuong* statistic is heavily biased towards the more heavily parameterised models. It chooses the ZIOP models all of the time when they are true, but fails to choose the true OP model with the bulk of the values falling in the indeterminate range. Of course, as with any Monte Carlo experiments, all of the above results are conditional on the specific experimental designs.

4. An application to tobacco consumption

Cigarette smoking has long been acknowledged as a public health issue. Yet a significant proportion of the population in both developed and developing countries smoke. Large amounts of public funds are spent worldwide on educational programs and promotional campaigns to reduce cigarette consumption. Empirical studies are crucial to help identify the socioeconomic and demographic factors associated with smoking, providing invaluable information to facilitate well-targeted public health policies.

4.1. The data

The data we use for the model are from the Australian National Drug Strategy Household Survey (NDSHS, 2001). In this data set, neither the monetary expenditures nor the physical quantities of tobacco consumed are reported. The information on individuals' consumption of tobacco is given via a discrete variable measuring the intensity of consumption. There have been seven surveys conducted through the NDSHS since 1985. The surveys collect information from individuals aged 14 and over on attitudes and consumption of several legal and illegal drugs. Measures have been put in place in the surveys to ensure confidentiality in order to reduce under reporting. In this paper, data from the three most recent surveys of 1995, 1998 and 2001 are used which involve a total of over 40,000 respondents. After removal of missing values, a sample of 28,813 individuals is used for estimation. This data set has been used in several previous studies (Cameron and Williams, 2001; Williams, 2003; Zhao and Harris, 2004).

Definitions of all variables used in the study are given in the Appendix. In particular, the information in the data concerning an individual's consumption of tobacco is collected through the question "*How often do you now smoke cigarettes, pipes or other tobacco products?*", where the responses take the form of one of the following choices: not at all ($y = 0$); smoking less frequently than daily ($y = 1$); smoking daily with less than 20 cigarettes per day ($y = 2$); and smoking daily with 20 or more cigarettes per day ($y = 3$). Table 6 presents some summary statistics on the observed smoking intensities. On average around 76% of individuals identify themselves as current non-smokers. With the way the survey questions are asked, these self-identified non-smokers will include genuine non-smokers, recent quitters, infrequent smokers who are not *currently* smoking, as well as potential smokers who might smoke when, say, the price falls. It could also be argued that these observations may include some misreporting respondents who prefer to identify themselves as non-smokers. The choices of consumption intensities are clearly ordered, thus presenting a good case for the ZIOP(C) model(s) in order to identify the different types of zero observations and their potentially different driving factors.

The participation decision of Eq. (1) is likely to be driven by factors relating to individuals' attitudes towards smoking and health concerns. Thus, r^* is likely to be related to the individuals' education levels and other standard socio-demographic variables such as income, marital status, age, gender and ethnic background that capture socioeconomic status. There are also studies in the literature suggesting a significant growth in the smoking prevalence of young females (Boreham and Shaw, 2002). To allow for the recent rise in participation rates among young females, a dummy variable for young females (defined as females under 25 years of age) is interacted with a time variable and included in $\mathbf{x}$ in Eq. (1).

In terms of the decision of the levels of consumption conditional on participation, economists have typically followed a standard consumer demand framework with special characteristics for addictive goods. Much work has been undertaken applying Becker and Murphy's (1988) theory of rational addiction in explaining consumer behaviour in terms of an individual's stock of addiction from past smoking (see Becker and Stigler, 1977; Chaloupka, 1991). Here, for the explanatory variables $\mathbf{z}$ in Eq. (3), we include standard demand-schedule variables such as income and own- and cross-drug prices. The related drug prices are included as there is evidence that certain drugs, in particular marijuana and alcohol, act as either compliments or substitutes to tobacco (see, for example, Cameron

Table 6
Summary of consumption frequencies

	1995		1998		2001		Combined	
	N	%	*N*	%	*N*	%	*N*	%
Tobacco								
Non-smoker	2644	**72.4**	7047	**72.1**	20113	**78.0**	29804	**76.0**
Weekly or less	120	**3.3**	504	**5.2**	937	**3.6**	1561	**4.0**
Daily, less than 20/day	600	**16.4**	1472	**15.1**	3351	**13.0**	5423	**13.8**
Daily, more than 20/day	286	**7.8**	749	**7.7**	1376	**5.3**	2411	**6.2**
Total	3650	**100**	9772	**100**	25777	**100**	39199	**100**

and Williams, 2001; Zhao and Harris, 2004). Data for marijuana prices were obtained from information provided by the Australian Bureau of Criminal Intelligence (ABCI, 2002) and the Australian Crime Commission (ACC, 2003). They are collected quarterly and are based on information supplied by covert police units and police informants. The consumer price indexes for tobacco and alcoholic drinks are obtained from the Australian Bureau of Statistics (ABS, 2003) for individual states. In addition, standard social demographic factors are also included in **z** to capture any heterogeneity in consumption behaviour among smokers.

Note that we allow the age factor to enter both equations. The participation decision is allowed to relate non-linearly to age by including age in natural logarithmic form. However, in the intensity of consumption equation, following a Becker and Murphy (1988) rational addiction approach, the likelihood that the age-consumption profile will be "n-shaped" is allowed for by including both linear and quadratic terms for age.

4.2. *The results*

In Table 7 we present some summary statistics from three models: an OP model conditional on **z** and treating all observed zeros indifferently; a ZIOP model conditional on both **x** and **z** that allows zero observations to come from two distinct sources; and a ZIOPC that further allows for correlation across the two error terms. Results for some ancillary parameters are also presented. As the magnitudes of the estimated coefficients of $\boldsymbol{\beta}$ and γ are somewhat meaningless, they are not presented here.[8]

The *LR* statistics clearly reject the OP model here, as does the *Hausman* one for the correlated version (the statistic for the uncorrelated version yielded a *negative* value). Furthermore, all of the information criteria, as well as the *Vuong* test, clearly suggest superiority of the ZIOP and ZIOPC models over the OP one. With the exception of *AIC*, the information criteria marginally favour the uncorrelated variant (ZIOP) to the correlated one (ZIOPC). Moreover, a *Wald* test on the estimated value of ρ also suggests that the correlation is not statistically significant.

The results are presented as marginal effects on the choice probabilities in Tables 8 and 9. For comparison purposes, we also include the results of a simple probit model to compare results on participation in Table 8. We only present those for the ZIOPC model as the ZIOP ones were very similar. Note that for variables appearing in both **x** and **z**, we have combined the two parts of the marginal effects, following Eq. (14). In Table 8, we present marginal effects on $\Pr(y=0)$ using a ZIOPC model and compare them with the results from the probit and OP models. For the ZIOPC model, we also decompose the overall marginal effect on $\Pr(y=0)$ into two parts: the effect on non-participation $\Pr(r=0)$ and the effect on participation with zero consumption $\Pr(r=1,\tilde{y}=0)$. In Table 9, we present marginal effects on the unconditional probabilities of all three positive levels of smoking ($y=1,2,3$), using an OP model *versus* the ZIOPC model.

[8] Details for the estimated coefficients are in Harris and Zhao (2004). Coefficients for several covariates exhibit opposite signs in $\boldsymbol{\beta}$ and γ: income has a positive coefficient in $\boldsymbol{\gamma}$ for conditional consumption but a negative one in $\boldsymbol{\beta}$ for participation, while the sole income coefficient in either a probit or an OP model is positive. Note also that, although the price variables are not included in **x** on *a priori* grounds, if included they were individually and jointly insignificant.

Table 7
Tobacco consumption: some estimated coefficients and summary statistics from three alternative models[a]

	OP		ZIOP		ZIOPC	
μ_1	0.155	(0.004)**	0.273	(0.011)**	0.272	(0.011)**
μ_2	0.920	(0.011)**	1.387	(0.031)**	1.383	(0.031)**
ρ					−0.068	(0.222)
$\ell(\boldsymbol{\theta})$	−21,995		−21,628		−21,626	
AIC	44,012		43,292		**43,291**	
LR versus OP			734**		738**	
Hausman versus OP			−689		1,259**	
BIC	44,206		**43,635**		43,643	
CAIC	44,227		**43,672**		43,681	
Vuong: versus OP			−13.830		−13.890	
Time	8.1		25.9		224.8	

[a] Standard errors are in parentheses, and (**) and (*) indicate statistical significance at 5% and 10% levels, respectively. Preferred model with regard to each information criteria is indicated with *bold*.

The marginal effects in Tables 8 and 9 highlight some interesting differences from alternative models for some explanatory factors, such as *Ln*(*Income*), *Pre-School*, *School*, *Young Female* and *Study*. A key example is the effect of income. This variable clearly acts as a social class proxy in the participation decision and, accordingly, one would expect *a priori* it to be positively associated with non-participation. Results based on the ZIOPC model suggest that a 10% increase in personal income results in a 0.0027 *rise* in the probability of non-participation, but a 0.0017 fall in the probability of participation with zero consumption. This latter effect indicates that tobacco is a *normal good* for participants. Overall, there is a 0.001 net positive effect on the probability of observing zero consumption for a 10% increase in personal income. However, basing policy advice on the probit (or OP) model results, one would conclude that income is positively related to participation as well as higher consumption.

Another example is the marginal effect of *Study*. Using simple probit and OP models, one would conclude that people who mainly study are *more* likely to be non-smokers (by 0.098 and 0.128, respectively, Table 8). With a single latent equation, we assume, in the case of OP, that there is a homogenous 'study' effect that affects an individual moving from non-smokers to smokers of higher levels ($y = 0, 1, 2, 3$) in the same direction. However, when a ZIOPC model is used, we assume that the observed smoking categories are the result of two distinct decisions of 'participation' and 'levels of consumption conditional on participation', on which *Study* can have opposite effects. Indeed, as shown in Table 8, the ZIOPC estimates that *Study* has a positive effect on participation decision but a negative effect on levels of consumption, leading to a statistically insignificant total effect of observing a zero ($y = 0$) outcome as the opposing effects cancel each other out. This contrasts the positive effects on non-participation by both the probit and OP models. In addition, the resulting marginal effects on the *unconditional* probabilities of levels of smoking ($\Pr(y = j), j = 1, 2, 3$) in ZIOPC is also the result of two sources: *ME*s on participation, $\Pr(r = 1)$, and *ME*s on levels of smoking conditional on participation, $\Pr(y = j|r = 1)$. For example, the −0.032 *ME* of *Study* on heavy smoking, $\Pr(y = 3)$, in Table 9 is the combined result of opposing effects

Table 8
Tobacco consumption: marginal effect for non-participation and zero consumption

			ZIOPC		
	Probit	OP	Non-participation	Zero consumption	Full
	$\Pr(y=0)$	$\Pr(y=0)$	$\Pr(r=0)$	$\Pr(r=1,\tilde{y}=0)$	$\Pr(y=0)$
Young female	0.021	–	−0.059	0.023	−0.036
	(0.004)**	–	(0.025)**	(0.010)**	(0.015)**
Actual age	0.006	0.004	0.015	−0.009	0.006
	(0.000)**	(0.000)**	(0.001)**	(0.001)**	(0.000)**
Ln (Income)	−0.014	−0.003	0.027	−0.017	0.010
	(0.004)**	(0.004)	(0.009)**	(0.006)**	(0.005)**
Male ×1	−0.018	−0.047	−0.095	0.023	−0.072
	(0.006)**	(0.005)**	(0.013)**	(0.009)**	(0.007)**
Married ×1	0.082	0.099	0.160	−0.039	0.121
	(0.006)**	(0.005)**	(0.013)**	(0.010)**	(0.007)**
Pre-school ×1	−0.012	0.009	0.054	−0.019	0.035
	(0.008)	(0.007)	(0.019)**	(0.014)	(0.009)**
Capital ×1	0.007	0.012	−0.008	0.020	0.012
	(0.006)	(0.005)**	(0.012)	(0.010)**	(0.007)*
Work ×1	−0.002	0.053	−0.008	0.049	0.041
	(0.008)	(0.008)**	(0.017)	(0.014)**	(0.009)**
Unemployed ×1	−0.129	−0.044	−0.061	0.004	−0.057
	(0.018)**	(0.014)**	(0.032)*	(0.022)	(0.018)**
Study ×1	0.098	0.128	−0.194	0.182	−0.012
	(0.010)**	(0.012)**	(0.061)**	(0.033)**	(0.032)
English ×1	−0.046	−0.057	−0.063	−0.004	−0.067
	(0.011)**	(0.012)**	(0.028)**	(0.021)	(0.014)**
Degree ×1	0.150	0.184	0.080	0.128	0.209
	(0.006)**	(0.008)**	(0.022)**	(0.019)**	(0.010)**
Diploma ×1	0.038	0.048	0.027	0.036	0.063
	(0.007)**	(0.007)**	(0.014)*	(0.012)**	(0.008)**
Year 12 ×1	0.056	0.069	0.020	0.060	0.079
	(0.007)**	(0.007)**	(0.018)	(0.014)**	(0.010)**
School ×1	0.174	0.170	−0.002	0.093	0.091
	(0.009)**	(0.019)**	(0.100)	(0.046)**	(0.058)
$Ln(P_{\mathrm{A}})$	–	0.322	–	0.300	0.300
	–	(0.078)**	–	(0.071)**	(0.071)**
$Ln(P_{\mathrm{M}})$	–	0.002	–	−0.004	−0.004
	–	(0.012)	–	(0.010)	(0.010)
$Ln(P_{\mathrm{T}})$	–	0.156	–	0.145	0.145
	–	(0.020)**	–	(0.019)**	(0.019)**

of *Study* on the two decisions. This contrasts the *ME* of −0.043 from an OP with only one source of impact.

4.3. Model evaluation

In Fig. 3 we present the observed sample proportions, average predicted probabilities and the probabilities evaluated at average covariates using a ZIOPC model for the four

Table 9
Tobacco consumption: marginal effects for non-zero consumption levels

	OP	ZIOPC	OP	ZIOPC	OP	ZIOPC
	Pr($y = 1$)	Pr($y = 1$)	Pr($y = 2$)	Pr($y = 2$)	Pr($y = 3$)	Pr($y = 3$)
Young female	–	0.006	–	0.021	–	0.008
	–	(0.003)**	–	(0.009)**	–	(0.003)**
Actual age	−0.001	−0.002	−0.002	−0.004	−0.001	−0.000
	(0.000)**	(0.000)**	(0.000)**	(0.000)**	(0.000)**	(0.000)**
Ln(income)	0.000	−0.003	0.002	−0.007	0.001	0.000
	(0.000)	(0.001)**	(0.002)	(0.003)**	(0.001)	(0.002)
Male ×1	0.006	0.010	0.026	0.042	0.016	0.021
	(0.001)**	(0.001)**	(0.003)**	(0.004)**	(0.002)**	(0.005)**
Married ×1	−0.012	−0.016	−0.054	−0.070	−0.033	−0.035
	(0.001)**	(0.002)**	(0.003)**	(0.005)**	(0.002)**	(0.005)**
Pre-school ×1	−0.001	−0.006	−0.005	−0.021	−0.003	−0.009
	(0.001)	(0.002)**	(0.004)	(0.006)**	(0.002)	(0.004)**
Capital ×1	−0.001	0.001	−0.007	−0.005	−0.004	−0.008
	(0.001)**	(0.001)	(0.003)**	(0.004)	(0.002)**	(0.003)**
Work ×1	−0.006	0.003	−0.029	−0.018	−0.018	−0.024
	(0.001)**	(0.002)	(0.004)**	(0.009)**	(0.003)**	(0.007)**
Unemployed ×1	0.005	0.006	0.024	0.032	0.015	0.020
	(0.002)**	(0.004)	(0.008)**	(0.011)**	(0.005)**	(0.008)**
Study ×1	−0.015	0.025	−0.070	0.021	−0.043	−0.032
	(0.002)**	(0.007)**	(0.007)**	(0.023)	(0.004)**	(0.012)**
English ×1	0.007	0.006	0.031	0.038	0.019	0.025
	(0.001)**	(0.003)*	(0.006)**	(0.009)**	(0.004)**	(0.007)**
Degree ×1	−0.022	−0.003	−0.101	−0.107	−0.061	−0.099
	(0.001)**	(0.002)	(0.005)**	(0.006)$^{*\ *}$	(0.003)**	(0.005)**
Diploma ×1	−0.006	−0.001	−0.026	−0.032	−0.016	−0.030
	(0.001)**	(0.002)	(0.004)**	(0.005)**	(0.002)**	(0.004)**
Year 12 ×1	−0.008	0.000	−0.038	−0.040	−0.023	−0.040
	(0.001)**	(0.002)	(0.004)**	(0.006)**	(0.002)**	(0.004)**
School ×1	−0.020	0.004	−0.093	−0.044	−0.057	−0.051
	(0.002)**	(0.011)	(0.011)**	(0.035)	(0.007)**	(0.013)**
$Ln(P_A)$	−0.038	0.011	−0.177	−0.144	−0.107	−0.166
	(0.009)**	(0.004)**	(0.043)**	(0.035)**	(0.026)**	(0.039)**
$Ln(P_M)$	−0.000	−0.000	−0.001	0.002	−0.001	0.002
	(0.001)	(0.000)	(0.006)	(0.005)	(0.004)	(0.006)
$Ln(P_T)$	−0.018	0.005	−0.085	−0.070	−0.052	−0.081
	(0.002)**	(0.002)**	(0.011)**	(0.009)**	(0.007)**	(0.010)**

smoking categories, which we denote zero, low, moderate and high. For Pr($y = 0$), we also present the probability of zeros arising from the regime of non-participation ($r = 0$) as the "selection component". As can be seen, the model fits the data well in terms of mimicking these sample proportions. Moreover, it is clear that the bulk of the probability mass of Pr($y = 0$) comes from non-participants with $r = 0$.

In Fig. 4 the overall age-smoking profile is plotted. The expected *n*-shaped profile is clearly evident for the smokers, and most pronounced for moderate and high levels of consumption. In terms of the probability of zero consumption, this finds a nadir at around the mid-late 20's, with the probability reaching nearly 0.9 at age 66.

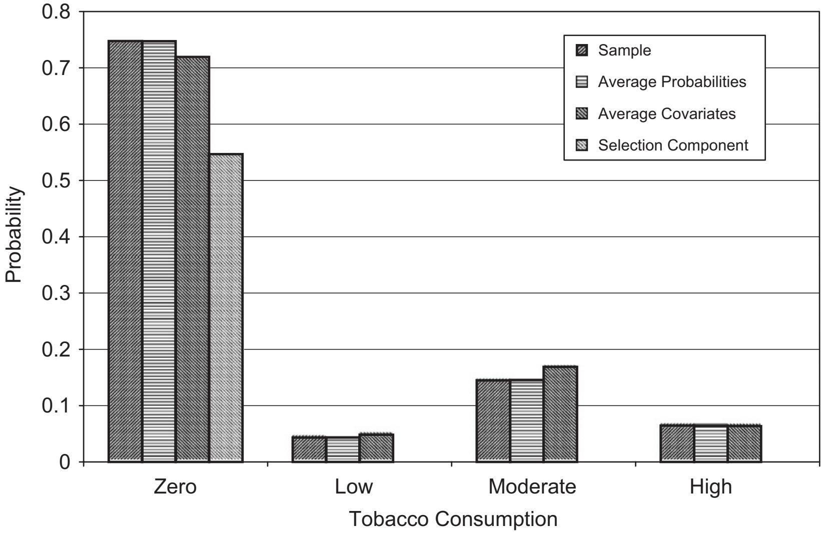

Fig. 3. Observed and predicted probabilities.

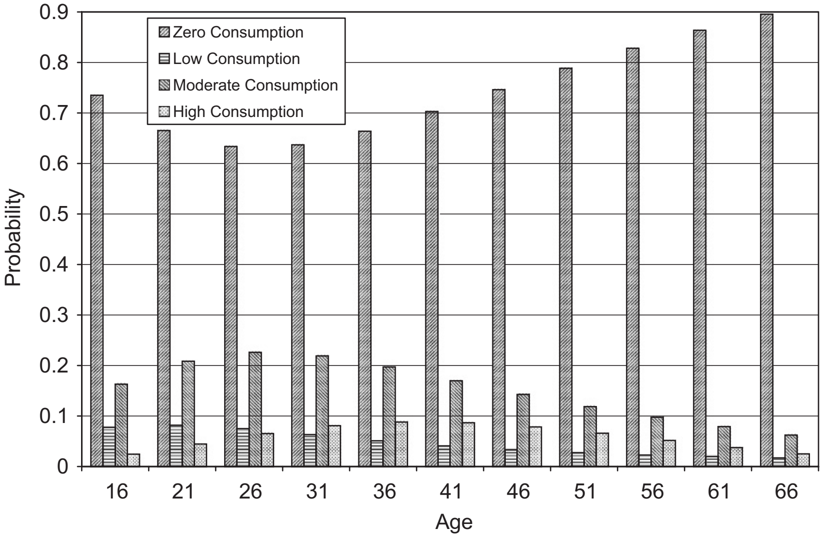

Fig. 4. Predicted probabilities by age.

As previously mentioned, one of the advantages of a ZIOPC is its ability to disentangle the total effect of a covariate on $\Pr(y=0)$ into those effects on the probabilities of the two types of zeros: $\Pr(r=0)$ and $\Pr(\widetilde{y}=0, r=1)$. This is illustrated in Fig. 5 for the effect of age. At younger ages, $\Pr(y=0)$ is dominated by potential consumers. However, as age increases and participation rates decline, the bulk of $\Pr(y=0)$ comes from genuine non-participation.

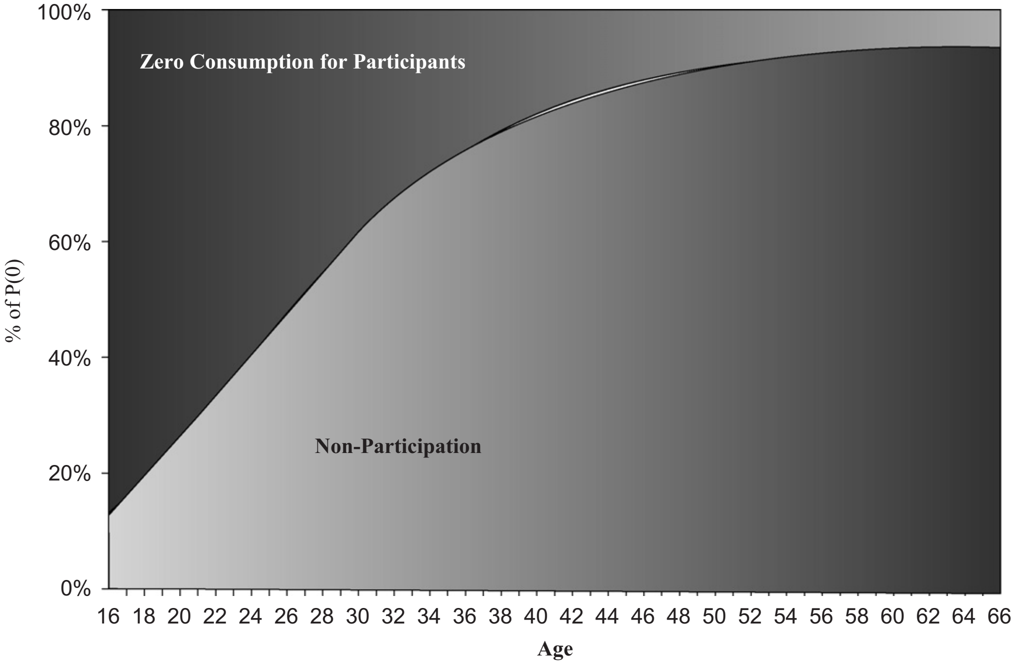

Fig. 5. Decomposition of $P(y = 0)$ into component parts.

While the results in Table 7 clearly suggest preference of the ZIOP(C) model(s) over the OP model, as a specification test, we also experimented with the more flexible models of multinomial logit (MNL) and multinomial probit (MNP). Without imposing restrictions in the correlation matrix, we encountered convergence problems with MNP using standard econometric software. The MNL model was found inferior on the basis of both *BIC* and *CAIC*. The log-likelihood values of the ZIOPC and MNL models were very close ($-21,626$ and $-21,601$, respectively), while the MNL has 20 additional parameters. In addition, *Hausman* tests clearly rejected the embodied *independence of irrelevant alternatives* property of the MNL model, and the smoking levels data here are quite clearly ordered in nature.

5. Conclusions

We propose a model for ordered discrete data that allows for the observed zero observations to be generated by two different behavioural regimes. Following double-hurdle and zero-inflated models, we extend the OP model to a zero-inflated OP model using a system of two latent equations with potentially different covariates. We also allow for the likely correlation between the two latent equations. The Monte Carlo experiments suggest that the model performs well in finite samples. Although not strictly valid, both the *LR* and *Hausman* statistics appear to provide useful general specification tests against the simpler OP model. The former may be preferred in terms of ease of computation and better finite sample performance. On the other hand, the *Vuong* test does not appear to have favourable small sample properties and tends to favour the ZIOP models, whilst information criteria seem to have good empirical properties with regard to selecting the correct model. If the *d.g.p.* is not zero split, the *LR* and *Hausman* statistics and the suggested model selection procedures based on information criteria should all correctly pick the OP model. However, even in this case, estimation of a ZIOP model will still provide accurate results in terms of the marginal effects, as the probability for Regime 1 tends to unity in the split decision such that ZIOP probabilities tend to OP ones. However, if the *d.g.p.* is zero split, the evidence is that the ZIOP models will be correctly selected if any of the *LR* and *Hausman* statistics, *Vuong* test or information criteria are used. With regard to differentiating between the ZIOP and ZIOPC, a *Wald* test of $\rho = 0$ would be preferred to the information based model selection procedures.

The models are applied to discrete data of tobacco consumption from a nationally representative Australian survey. The empirical application demonstrates the advantages of the ZIOP(C) model in separating the different behavioural schemes for participants and non-participants. In particular, we allow for the split of the observed non-users (“zeros”) into two groups: those of non-participants who choose not to smoke due to health concerns or other non-economic factors, and those zero consumption potential users who may be the result of a demand-schedule corner solution and are therefore responsive to economic factors such as prices and income. The example shows that the use of a conventional OP model would confuse the effects of some important explanatory variables that have opposing impacts on the two schemes.

The ZIOP(C) model has important advantages over the conventional OP model. It can be used to estimate the proportion of zeros coming from each regime, and how this split changes with observed characteristics. The proposed model allows for the identification of variables that are important in each regime. This is potentially very important for policy analysis.

Acknowledgements

We wish to thank John Gweke, three anonymous referees and one anonymous Associate Editor, as well as William Greene, Don Poskitt, Max King, Rob Hyndman, Tim Fry and Brett Inder for constructive comments. The usual caveats apply. Excellent research assistance was provided by Preety Ramful and Xiaohui Zhang. Funding from the Australian Research Council is kindly acknowledged by both authors.

Appendix A. Definition of variables

y	Levels of tobacco consumption; $y = 0$ if not current smoker, $y = 1$ if smoking weekly or less, $y = 2$ if smoking daily with less than 20 cigarettes per day, and $y = 3$ if smoking daily with 20 or more cigarettes per day.
Ln(Age)	Logarithm of actual age.
Age	Actual age divided by 10.
Age square	*Age* squared and divided by 10.
Male	1 for male and 0 for female.
Married	1 if married or *de facto* and 0 otherwise.

Pre-school	1 if the respondent has pre-school aged child/children and 0 otherwise.
Capital	1 if the respondent resides in a capital city, and 0 otherwise.
Work	1 if mainly employed and 0 otherwise.
Unemployed	1 if unemployed and 0 otherwise.
Study	1 if mainly study and 0 otherwise.
Other	1 if retired, home duty, or volunteer work and 0 otherwise. This variable is used as the base of comparison for work status dummies and is dropped in the estimation.
English	1 if English is the main language spoken at home for the respondent and 0 otherwise.
Degree	1 if the highest qualification is a tertiary degree and 0 otherwise.
Diploma	1 if the highest qualification is a non-tertiary diploma or trade certificate, and 0 otherwise.
Year 12	1 if the highest qualification is Year 12 and 0 otherwise.
School	1 if still studying in school and 0 otherwise.
Noqual	1 if the highest qualification is below Year 12 and 0 otherwise. This variable is used as the base of comparison for education dummies and is dropped in the estimation.
$Ln(P_{\mathrm{T}})$	Logarithm of real price index for tobacco, divided by 10.
$Ln(P_{\mathrm{A}})$	Logarithm of real price index for alcoholic drinks, divided by 10.
$Ln(P_{\mathrm{M}})$	Logarithm of real price for marijuana measured in dollars per ounce, divided by 10.
Ln(Income)	Logarithm of real personal annual income before tax measured in thousands of Australian dollars, divided by 10.
Young female	A binary dummy for female aged 25 years or younger, interacted with an annual time trend $t = 1, 2, 3$.

References

ABCI, 2002. Australian Illicit Drug Report. Australian Bureau of Criminal Intelligence.

ABS, 2003. Consumer Price Index 14th Series: by region by group, sub-group and expenditure class—alcohol and tobacco, Cat. No. 6455.0.40.001, Australian Bureau of Statistics.

ACC, 2003. Australian Illicit Drug Report 2001-02, Australian Crime Commission.

Andrews, D., Ploberger, W., 1995. Admissibility of the likelihood ratio test when a nuisance parameter is present only under the alternative. The Annals of Statistics 23 (5), 1609–1629.

Becker, G., Murphy, K., 1988. A theory of rational addiction. Journal of Political Economy 96 (4), 675–700.

Becker, G., Stigler, G., 1977. De gustibus non est disputandum. American Economic Review 68 (1), 76–90.

Boreham, R., Shaw, A., 2002. Drugs Use, Smoking and Drinking Among Young People in England in 2001. The Stationary Office.

Cameron, C., Trivedi, P., 1998. Regression Analysis of Count Data. Econometric Society Monographs. Cambridge University Press, Cambridge, UK.

Cameron, L., Williams, J., 2001. Cannabis alcohol and cigarettes: substitutes or compliments. The Economic Record 77 (236), 19–34.

Chaloupka, F., 1991. Rational addictive behaviour and cigarette smoking. Journal of Political Economy 99 (4), 722–742.

Chesher, A., Smith, R., 1997. Likelihood ratio specification tests. Econometrica 65 (3), 627–646.

Cragg, J., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39, 829–844.

Greene, W., 1994. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC-94-10, Stern School of Business, New York University, Stern School of Business, New York University.

Greene, W., 2003. Econometric Analysis, fifth ed. Prentice-Hall, Englewood Cliffs, NJ, USA.

Harris, M., Zhao, X., 2004. Modelling tobacco consumption with a zero-inflated ordered probit model. Working Paper 14/04, Monash University, Department of Econometrics and Business Statistics, Monash University, Australia.

Hausman, J., 1978. Specification tests in econometrics. Econometrica 46, 1251–1271.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Heilbron, D., 1989. Generalized linear models for altered zero probabilities and overdispersion in count data. Discussion paper, University of California, University of California, San Francisco.

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

Maddala, G.S., 1983. Limited Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge, UK.

Marcus, A., Greene, W., 1985. The determinants of rating assignment and performance. Discussion Paper Working Paper CRC528, Center for Naval Analyses.

Mullahey, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics 33, 341–365.

Mullahey, J., 1997. Heterogeneity, excess zeros and the structure of count data models. Journal of Applied Econometrics 12, 337–350.

NDSHS, 2001. Computer Files for the Unit Record Data from the National Drug Strategy Household Surveys.

Pohlmeier, W., Ulrich, V., 1995. An econometric model of the two-part decision-making process in the demand for health care. Journal of Human Resources 30, 339–361.

Smith, M., 2003. On dependency in double-hurdle models. Statistical Papers 44, 581–595.

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–334.

Williams, J., 2003. The effects of price and policy on marijuana use: what can be learned from the australian experience. Health Economics 13 (2), 123–137.

McKelvey, R., Zavoina, W., 1975. A statistical model for the analysis of ordinal level dependent variables. Journal of Mathematical Sociology 4, 103–120.

Zhao, X., Harris, M., 2004. Demand for marijuana, alcohol and tobacco: participation frequency, and cross-equation correlation. Economic Record 80 (251), 394–410.