Zero-inflated ordinal data models with application to sport (*in*)activity

Modelli per dati ordinali zero-inflazionati con applicazione all'(in)attività sportiva

Maria Iannario and Rosaria Simone

Abstract Traditional models for ordinal data (as CUB models or cumulative models with logit/probit link, among others) present limits in explaining the surplus of zero observations, especially when the zeros may relate to two distinct situations of non-participation/inactivity and infrequent participation, for instance. We consider an extension of standard models: zero-inflated CUB models and zero inflated ordered cumulative (ZIOC) probit/logit models handling the GECUB models and using a double-hurdle combination of a split (logit/probit) model and an ordered probit/logit model, respectively. Both extensions, potentially, relate to different sets of covariates. Finally, models are applied to Sport surveys. Specifically the paper investigates the determinants of sport (*in*)activity: the frequency and the probability of sports participation. It distinguishes between genuine "non-participants" and the ones who do not participate at a time but might do under different circumstances.

Abstract *I modelli tradizionali per i dati ordinali (come modelli* CUB *o cumulativi con link logit/probit, tra gli altri) presentano limiti nello spiegare il surplus di osservazioni nella categoria zero, specialmente quando gli zeri possono riguardare due distinte situazioni di non partecipazione/non attività e/o partecipazione non frequente. Il lavoro propone un'estensione di modelli standard: i modelli* CUB zero inflated *e i modelli ordinal probit/logit con inflazione di zeri (*ZIOC*). I primi costituiscono una revisione dei modelli* GECUB *(modelli* CUB *con effetto shelter), i secondi costituiscono una mistura di modelli (probit/logit) dicotomici e modelli probit/logit ordinali. Entrambe le estensioni possono riferirsi a diversi gruppi di covariate. Infine, i modelli sono applicati a dati rilevati da indagini sullo sport. In particolare, lo studio esplora le determinanti dell'(in)attività sportiva: la frequenza e la probabilità di partecipazione ad attività sportive. Distingue tra veri "non partecipanti"*

Maria Iannario
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22 - Napoli
e-mail: maria.iannaro@unina.it

Rosaria Simone
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22 - Napoli
e-mail: rosaria.simone@unina.it

e coloro che non partecipano al momento dell'indagine, ma potrebbero se in circostanze diverse.

Key words: CUB models, Ordinal logit/probit models, Ordered outcomes, discrete data, sport inactivity, zero-inflated responses

1 Introduction

Excess of zeros is a commonly encountered phenomenon that limits the use of traditional regression models for analysing ordinal data in contexts where respondents express a graduated perception on a specific item or experiments identify levels of increasing assessments.

The situation occurs with ordinal scales in which there is an anchor that represents the absence of the symptom or activity, such as none, never or normal. This level usually tagged *zero* may be scored by respondents certainly not at risk (without symptom or who do not practice any activity/exercise) and respondents with a non-zero probability of risk.

Survey data concerning epidemiological studies or choices, particularly those that refer to an explicit time dimension, may include genuine non-participants whatever the circumstances are, as well as individuals who would decide to participate if the circumstances were different. It is, therefore, likely that these two types of zeros are driven by different behaviour. One example is a study by Harris and Zhao (2007) on the consumer choice problem of tobacco consumption or the analysis of Downward et al. (2011) on sports participation, among others.

Aim of the paper is introducing methodologies that allow users of ordinal scale data to more accurately model the distribution of ordinal outcomes in which some subjects are susceptible to exhibit the response and some are not (i.e. the dependent variable exhibits zero inflation). The study explores the determinants of sport (*in*)activity: the frequency and the probability of sports participation. It distinguishes between genuine "non-participants" and the ones who do not participate at a time but might do under different circumstances. Thus, it includes whether or not to participate in sport and, subsequently, what intensity of participation is undertaken. It is able to distinguish between structured and sampling zeros implementing some results obtained for count data in the ordinal data context.

With respect to the standard models for ordinal data the new methodologies exceed some gaps related to the model of zeros by taking into account the potentially two-fold decision made with respect to participation. Here we propose extensions of standard models: zero-inflated CUB (ZICUB) models and zero inflated ordered cumulative (ZIOC) probit/logit models handling the GECUB models and using a double-hurdle combination of a split (logit/probit) model and an ordered probit/logit models, respectively. Both extensions, potentially, relate to different sets of covariates. The modelling assumption is that different decisions govern the choice to participate and the frequency of participation in sport. The remainder of the paper is as follows.

Section 2 reviews the methods used for the analysis. Section 3 describes the data set and main estimation results with a summary of the main findings and opportunities for further research.

2 Methods

Let Y be a discrete random variable that assumes the ordered values of $0,1,...,J$. Standard ordinal cumulative (Agresti, 2010) or CUB models (Piccolo, 2003) map a single latent variable Y^* to the observable Y, with Y^* related to a set of covariates or consider the response as a weighted mixture of respondents' propensity to adhere to a meditated choice (formally described by a shifted Binomial random variable) and a totally uninformative choice (described by a discrete Uniform distribution) with a possible *shelter* effect (Iannario, 2012; Iannario and Piccolo, 2016), respectively. Here we propose a zero inflated cumulative (ZIOC) model that involves two latent equations with uncorrelated error terms: a logit/probit equation and an ordered logit/probit equation by introducing ZIOL/ZIOP models (subsection 2.1). Or in order to further disentangle the *inflated effect* concentrated at category *zero* we may introduce a variant of GECUB models (subsection 2.2).

2.1 Zero Inflated Cumulative Models

Let r denote a binary variable indicating the split between Regime 0 ($r=0$, for "non participants") and Regime 1 ($r=1$ for "participants"), which is related to the latent variable $r^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$ where $\boldsymbol{x}$ is a vector of p individual characteristics (covariates) that determine the choice of regimes, $\boldsymbol{\beta}$ is a p-vector of unknown regression parameters, and ε is a random variable with cumulative distribution function $G_\varepsilon(.)$. Accordingly, the probability of an individual being in Regime 1 is given by

$$Pr(r=1|\boldsymbol{x}) = Pr(r^*>0|\boldsymbol{x}) = G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta}),$$

where we assume $G_\varepsilon(.)$ strictly increasing and symmetric around zero. Standard choices for the distribution function are the logit link function, $G(t) = 1/(1+e^{-t})$, corresponding to the logistic distribution, or the probit link function, $G(t) = \Phi(t)$, with Φ the cdf of the standard normal distribution.

Conditional on $r=1$, respondents levels under Regime 1 are represented by $\tilde{Y}(\tilde{Y}=0,1,\ldots,J)$, which is generated by a cumulative link model based upon a second underlying latent variable $\tilde{Y}^*$, where

$$\tilde{Y}^* = \boldsymbol{z}'\boldsymbol{\gamma} + u,$$

with $\boldsymbol{z}$ being a vector of covariates with unknown parameters $\boldsymbol{\gamma}$ and $u \sim G_\varepsilon(.)$. The observed ordinal variable $\tilde{Y}$ takes as values the labels 0 if $\tilde{Y}^* \leq 0$, J if $\tilde{Y}^* \geq \alpha_{J-1}$

otherwise,

$$\tilde{Y} = j \qquad \Longleftrightarrow \qquad \alpha_{j-1} < \tilde{Y}^* \leq \alpha_j \qquad j = 1,2,\ldots,J-1\,, \qquad j > 2\,,$$

where $\alpha_j\,(j = 1,\ldots,J-1)$ are the intercept values to be estimated in addition to the covariate coefficients $\boldsymbol{\gamma}$. Notice that Regime 1 also allows for zero scores. That is, to observe $Y = 0$ 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 score, instead, we require jointly that the individual is a participant $(r = 1)$ and $\tilde{Y}^* > 0$. If we assume that the error terms from the first stage equation and the second stage cumulative outcome equation, that is e and u, are not correlated the probability mass function of the ZIOC model is

$$P_r(Y) = \begin{cases} P_r(Y = 0 \mid \boldsymbol{z},\boldsymbol{x}) = P_r(r = 0 \mid \boldsymbol{x}) + P_r(r = 1 \mid \boldsymbol{x})\, P_r\left(\tilde{Y} = 0 \mid \boldsymbol{z}, r = 1\right) \\ P_r(Y = j \mid \boldsymbol{z},\boldsymbol{x}) = P_r(r = 1 \mid \boldsymbol{x})\, P_r\left(\tilde{Y} = j \mid \boldsymbol{z}, r = 1\right)\ (j = 1,\ldots,J) \end{cases}$$

$$= \begin{cases} P_r(Y = 0 \mid \boldsymbol{z},\boldsymbol{x}) = [1 - G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta})] + G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta}) G_\varepsilon(-\boldsymbol{z}'\boldsymbol{\gamma}) \\ P_r(Y = j \mid \boldsymbol{z},\boldsymbol{x}) = G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta})[G_\varepsilon(\alpha_j - \boldsymbol{z}'\boldsymbol{\gamma}) - G_\varepsilon(\alpha_{j-1} - \boldsymbol{z}'\boldsymbol{\gamma})]\,(j = 1,\ldots,J-1) \\ P_r(Y = J \mid \boldsymbol{z},\boldsymbol{x}) = G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta})[1 - G_\varepsilon(\alpha_{J-1} - \boldsymbol{z}'\boldsymbol{\gamma}). \end{cases}$$

In this way, the probability of a zero score has been inflated as it is a combination of the probability of zero consumption from the cumulative model framework and the probability of non-participation from the split logit/probit model. Notice that the choice of distribution function G_ε allows to consider the Zero Inflated Ordinal Probit (ZIOP) as in Harris and Zao (2007) or Zero Inflated Ordinal Logit (ZIOL) models. Once the full set of probabilities has been specified, and given an *iid* sample $(i = 1,...,n)$ from the population on $(Y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, the parameters of the full model $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\alpha}')'$ may be estimated using the maximum likelihood (ML) methods. The log-likelihood function is $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=0}^{J} I[Y_i = j] \log P_r(Y_i = j|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})$, where $I[Y_i = j]$ is the indicator function of $(Y_i = j)$.

2.2 Zero Inflated CUB Models

Let $\breve{Y} \sim CUB_{she=1}(\breve{\pi}, \xi, \delta; J+1)$ be a CUB distributed random variable over $J+1$ categories and *shelter* at $c = 1$:

$$P_r\left(\breve{Y} = j \mid \breve{\pi}, \xi, \delta\right) = \delta\, D_j^{(c)} + (1-\delta)\left[\breve{\pi}\, b_j(\xi) + (1-\breve{\pi})\, h_j\right],\ j = 1,2,\ldots,J+1,$$

where $h_j = \dfrac{1}{J+1}$ is the discrete Uniform distribution over the given support and $b_j(\xi)$ denotes the shifted Binomial distribution with parameter $1-\xi$.

Then, a ZICUB model for the response variable $Y \in \{0,\ldots,J\}$ is specified by setting $Y = \breve{Y} - 1$. In this way

$$Pr(Y) = \begin{cases} Pr(Y=0|\boldsymbol{\theta}) = \delta + (1-\delta)\left[\pi b_1(\xi_i) + (1-\pi)\dfrac{1}{J+1}\right] \\ Pr(Y=j|\boldsymbol{\theta}) = (1-\delta)\left[\pi b_{j+1}(\xi) + (1-\pi)\dfrac{1}{J+1}\right], \quad j=1,\ldots,J. \end{cases}$$

In addition to examine the effects of risk factors on the response variable it may be proposed the inclusion of covariates on the parameters through canonical logit link:

$$logit(\delta_i) = \boldsymbol{\omega}'\boldsymbol{x}_i;\ logit(\pi_i) = \boldsymbol{\eta}'\boldsymbol{z}_i;\ logit(\xi_i) = \boldsymbol{\zeta}'\boldsymbol{w}_i\,.$$

Here, $\boldsymbol{\theta} = (\boldsymbol{\omega}', \boldsymbol{\eta}', \boldsymbol{\zeta}')'$ is the parameter vector characterizing the distribution of $(Y_1, Y_2, \ldots, Y_n)$ with $\boldsymbol{\omega}', \boldsymbol{\eta}', \boldsymbol{\zeta}'$ denoting the parameter vector for the *shelter*, uncertainty and feeling components, respectively, and $\boldsymbol{x}_i \in \boldsymbol{X}, \boldsymbol{z}_i \in \boldsymbol{Z}$ and $\boldsymbol{w}_i \in \boldsymbol{W}$ being the selected covariates for the i-*th* subject of the three components. The zero-inflated variant of GECUB models also assumes that some zeros are observed due to a specific structure in the data.

Here, given an observed random sample $(Y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{w}_i)$, for $i = 1, 2, \ldots, n$, the log-likelihood function is $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=0}^{J} I[y_i = j] \log Pr(Y_i = j|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{w}_i, \boldsymbol{\theta})$, where $I[Y_i = j]$ is the indicator function of $(Y_i = j)$.

3 Data and application

The determinant of sport (*in*)activity will be discussed on the basis of two case studies involving data collected in 2016 and 2017, respectively, through a web link related to the BDsports project (http://bodai.unibs.it/bdsports/). The case studies have been selected to highlight pitfalls and advantages of the two main proposals and to allow the distinction between genuine inactive respondents and the ones who do not play sport at a time of surveys.

In the last decade the modelling of sports participation decision has increased in complexity. The sports participation variable is measured in different ways; relatively few studies consider the time spent on sports participation or the frequency of such participation as we done in this paper. The dependent variable under investigation for 2016 is a rating on a 7 point scale whereas for 2017 is a rating on 11 categories (see Figure 1), asking each respondent the time dedicated to sport practice on weekly basis: from '0 = Rarely practiced any sport/not practiced any sport at all', '1 = Less than one hour' up to '6 = More than 7 hours' (up to '10' for 2017). Notice that the two surveys are about two different main topics (sport preferences and habits for the first survey, on the exercise addiction for the second ones); however both of them present a rating question on sport activity. Because the dependent variable is ordered rather than continuous and because, as noted in the Introduction, 'zero' participation could measure never participated (genuine inactive) or not recently/rarely participated a Zero-inflated ordered (ZIOP) estimator and ZICUB models are employed for 2016 whereas a ZIOL and ZICUB models for 2017. The modelling assumption is that different decisions govern the choice to partici-

pate and the frequency of participation in sport. Hurdle models are not considered by following what it is in Downward et al. (2011).

Evidence in the literature reveals that the probability of sports activity decreases with *age* (Barber and Havitz, 2001; Downward and Rasciute, 2010; among others) with less difference in gender among the older adults (Bauman et al. 2009). *Gender*, in fact, is the other covariate that has a highly important influence on sports activity. There is evidence about the fact that men, in general, not only participate in sport more than women (Downward and Rasciute, 2010; Eberth and Smith, 2010; Hovemann and Wicker, 2009; Lera-López and Rapún-Gárate, 2007) but they also show a higher frequency of participation (Barber and Havitz, 2001; Eberth and Smith, 2010). These differences may be attributed to biological factors, and cultural and social influences (Humphreys and Ruseski, 2010). Another determinant of sport (*in*)activity is the smoking habit; it has (with alcohol consumption) negative effects on sport practice especially in relation to age (Perretti et al. 2002).

Thus, in our analysis the selected covariates for 2016 are *gender, age* and the *smoking* habit. Results concerning a sample of $n = 647$ respondents are in Table 1 with the thresholds (cutpoints) on the underlying scale for ZIOP model $\hat{\alpha}_1 = -4.683$, $\hat{\alpha}_2 = -0.918$, $\hat{\alpha}_3 = -0.751$, $\hat{\alpha}_4 = -0.188$, $\hat{\alpha}_5 = -0.026$, $\hat{\alpha}_6 = 1.382$.

As revealed by estimation results of ZIOP model (Table 1) and mentioned in the literature sport activity reduces with older age; furthermore, smokers are generally inactive as well as women. Both estimated models confirm the effect of *age* for the inflation in the zero category that in ZICUB models is explicitly due to *smoking* habits. The best performance in terms of BIC index is for ZICUB model (bold in Table).

Similar results have been obtained for the analysis of the second survey (2017) where *gender, age* and the dichotomous response to the question "do you practice any sport or physical activity?" are considered ($n = 554$). Results are in Table 2 with the thresholds (cutpoints) on the underlying scale for ZIOL model $\hat{\alpha}_1 = -6.975$,

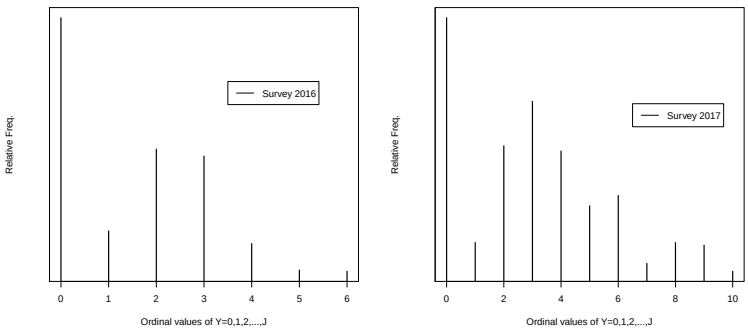

Fig. 1 Frequency distribution of the time spent on sports participation ($J = 0, 1, \ldots, 6$; Survey 2016 left side) ($j = 0, 1, \ldots, 10$; Survey 2017 right side)

Table 1 Regression results for ZICUB and ZIOP models

Models	Covariates	Parameters	Estimates	StdErr
ZICUB	*Constant*	$\hat{\omega}_0$	−0.871	0.115
	Smoke	$\hat{\omega}_1$	0.556	0.201
	Constant	$\hat{\eta}_0$	2.773	0.715
	Age	$\hat{\eta}_1$	−0.057	0.022
	Constant	$\hat{\zeta}_0$	0.283	0.063
	Woman	$\hat{\zeta}_1$	0.163	0.085
	$\ell(\boldsymbol{\theta})$	−1056.233	*BIC*	**2172.876**
ZIOP	Age	$\hat{\beta}_1$	0.611	0.087
	Constant	$\hat{\gamma}_0$	−0.309	0.112
	Woman	$\hat{\gamma}_1$	−0.345	0.104
	Smoke	$\hat{\gamma}_2$	0.014	0.005
	$\ell(\boldsymbol{\theta})$	−1067.021	*BIC*	2177.189

$\hat{\alpha}_2 = -1.128$, $\hat{\alpha}_3 = -0.825$, $\hat{\alpha}_4 = -0.571$, $\hat{\alpha}_5 = -0.543$, $\hat{\alpha}_6 = -0.074$, $\hat{\alpha}_7 = 0.011$, $\hat{\alpha}_8 = 0.300$, $\hat{\alpha}_9 = 1.034$.

Here it possible to notice the different impact of *age* on the uncertainty component for ZICUB model; for ZIOL model, instead, it has been confirmed the increasing inactivity for older respondents. Furthermore, to be woman and the answer to no sport/physical activity practiced represent requisites which express sport inactivity. In ZICUB model the effect of *gender* influences the feeling component by explaining woman inactivity, especially for "no practice at all" respondents. Generally this last model presents a better performance (BIC index in bold); here the inflation in zero consistently increases with the response "no practice". Data and the *R* code for the implementation of the methods are available upon request from Authors.

Finally, by testing different covariates and models to explain sport (*in*)activity it turns out that *age, gender, smoking* habit and *no sport/physical activity* exercised affect the occurrence of sedentary behaviour: given all these drivers it is possible to analyse the effect of age for zero inflation in ZIOC models, and smoking habits and no sport/physical activity practised for ZICUB models. Both implemented methods confirm the main results of the literature. Although the choice between the two zero-inflated approaches is generally based on the aim of the study, the evaluation in terms of fitting results and the interpretation of covariates may address the selection. Moreover, it is important to highlight some computational drawbacks related to the performance of ZIOC models.

Generally assessing the nature of the zero scores is becoming a more and more relevant issue demanding for the use of both the proposals. They can be used to estimate the proportion of zeros coming from each regime, and to evaluate how the split changes with observed characteristics.

Simulation studies will be planned to further validate and compare the efficacy of the proposals.

Table 2 Regression results for ZICUB and ZIOL models

Models	Covariates	Parameters	Estimates	Std Err
ZICUB	*Constant*	$\hat{\omega}_0$	−5.281	2.357
	No practice at all	$\hat{\omega}_1$	7.202	2.382
	Constant	$\hat{\eta}_0$	−2.895	0.860
	Age	$\hat{\eta}_1$	0.112	0.035
	Constant	$\hat{\zeta}_0$	0.347	0.130
	Woman	$\hat{\zeta}_1$	1.175	0.282
	No practice at all	$\hat{\zeta}_2$	0.212	0.084
	$\ell(\boldsymbol{\theta})$	−1085.695	*BIC*	**2215.61**
ZIOL	Age	$\hat{\beta}_1$	3.595	0.764
	Constant	$\hat{\gamma}_0$	−0.429	0.398
	Woman	$\hat{\gamma}_1$	−5.177	0.529
	No practice at all	$\hat{\gamma}_2$	−0.060	0.012
	$\ell(\boldsymbol{\theta})$	−1078.085	*BIC*	2244.611

References

1. Agresti A.: *Analysis of Ordinal Categorical Data*, 2nd Ed., J.Wiley & Sons, Hoboken (2010).
2. Barber, N., Havitz, M.E.: Canadian participation rates in ten sport and fitness activities. *Journal of Sport Management*, **15**, 51–76 (2001).
3. Bauman, A., Sallis, J., Dzewaltowski, D., Owen, N.: Toward a better understanding of the influences on physical activity. *American Journal of Preventive Medicine*, **23** (2S), 5–14 (2002).
4. Downward, P., Lera-López, F., Rasciute, S.: The Zero-Inflated ordered probit approach to modelling sports participation, *Economic Modelling*, **28**, 2469-2477 (2011).
5. Downward, P., Rasciute, S.: The relative demands for sports and leisure in England. *European Sport Management Quarterly*, **10** (2), 189–214 (2010).
6. Eberth, B., Smith, M.: Modelling the participation decision and duration of sporting activity in Scotland. *Economic Modelling*, **27** (4), 822–834 (2010).
7. Harris, N.M., Zhao, X.: A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *Journal of Econometrics*, **141** (2), 1073–1099 (2007).
8. Hovemann, G., Wicker, P.: Determinants of sport participation in the European Union. *European Journal for Sport and Society*, **6** (1), 51–59 (2009).
9. Humphreys, B., Ruseski, J.E.: The economic choice of participation and time spent in physical activity and sport in Canada, *Working Paper* No 201014. Department of Economics, University of Alberta (2010).
10. Iannario, M.: Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications*, **21**, 1–22 (2012).
11. Iannario, M., Piccolo, D.: A generalized framework for modelling ordinal data. *Statistical Methods and Applications*, **25**, 163–189 (2016).
12. Lera-López, F., Rapún-Gárate, M.: The demand for sport: sport consumption and participation models. *Journal of Sport Management*, **21**, 103–122 (2007).
13. Peretti-Watel P., Beck, F., Legleye, S.: Beyond the U-curve: the relationship between sport and alcohol, cigarette and cannabis use in adolescents. *Addiction*, **97**, 707–716 (2002).
14. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85–104 (2003).