Count data regression: modelling diversification in sports participation in Spain

Jaume García, *Universitat Pompeu Fabra, Barcelona (Spain)*
Cristina Muñiz, *Universidad de Oviedo (Spain)*
María José Suárez, *Universidad de Oviedo (Spain)*

Abstract

Count data models are specifically designed to deal with those cases where the dependent variable is an integer non-negative variable, taking a small number of (low) values, which is the usual situation when the variable to be explained represents the number of times a particular event occurs. This chapter presents an overview of the specific features of the count data models most commonly used in the economic literature, paying special attention to how the zeros are generated. An empirical illustration from the sports economics literature is also provided. Using data from the *Spanish Survey of Sporting Habits (2020)*, individual diversification of sports activity, measured by the number of sports practised during a year, is studied. The empirical analysis has been implemented using *Stata* software. It takes into account the specific features of the dependent variable by estimating different count data models, starting with the standard versions used in the microeconometric literature (Poisson and Negative Binomial models) and extending these basic models by considering different specifications in terms of how the zeros are generated. Finally, specific attention is devoted to the interpretation of the estimated coefficients and the calculation of the marginal effects.

Keywords: overdispersion, count variables, excess of zeros, zero-inflated models, hurdle models

Corresponding author: Jaume Garcia, jaume.garcia@upf.edu

1. Introduction

Count data models are a type of discrete choice models characterized by the fact that the dependent variable is the count of the number of events that have occurred in a given interval. Examples of count variables are the number of visits to a doctor (Pohlmeier and Ulrich 1995), the number of inpatient stays in a hospital (Geil 1997), the number of goals scored in a football game (Inan 2021), the number of cultural events attended in a certain period (Muñiz et al. 2014), the number of trips (Jang 2005), the number of patents applied for by a firm (Hausman et al. 1984) or the number of mergers and acquisitions made (Agrawal and Sensarma 2007). In all cases, the outcome only takes integer and non-negative values. Moreover, it usually takes a small number of values, including zero, and its distribution is right-skewed.

The statistical analysis of count variables goes back a long way in time. Cameron and Trivedi (2014) point out that the Poisson distribution, which is the simplest distribution that takes into account the specific characteristics of the count variables, was first proposed in 1837 by Siméon Denis Poisson, but there is an open debate about whether it was Abraham de Moivre who first referred to that distribution. Some applications were already made in the 19th century, while the negative binomial distribution was derived in the 1920s. However, the development of more advanced specifications and the generalization of the use of these models in econometrics occurred mainly in the second half of the 20th century.

Several surveys, chapters and books about count data models have been published in recent decades (e.g. Cameron and Trivedi 1986 2014; Winkelmann 2008; Hilbe 2014; Dupuy 2018; Tang et al. 2023). The objective of this chapter is to present the main specifications applied in economic analysis when using cross-section data,[1] focusing on how the zeros are modelled, given that in most economic studies the outcome takes the value zero for a significant proportion of the sample. Depending on the reasons for observing a zero, some models may be more suitable than others.

In particular, this review starts with the simplest specifications, the Poisson and the Negative Binomial distributions. These specifications do not deal with the excess of zeros problem, so the zero-inflated versions of the Poisson and the Negative Binomial are

[1] A discussion of the issues associated with the estimation of panel data models for count data can be found in Cameron and Trivedi (2005, 2022) and Sun and Zhao (2013).

explained. Then, an alternative specification that has also been applied in the literature is introduced, the hurdle model, highlighting the main differences in the specification of zeros between this model and the zero-inflated version. Finally, ordered models –intended for ordered qualitative outcomes- are proposed as an option that could also be used for count variables.

Additionally, the implementation of the models is illustrated with an empirical example from sports economics. Specifically, individual sports participation is studied, measured by the number of sports the individual has practised during the previous year. This variable can be related to one of the four dimensions of the FITT principles (frequency, intensity, time and type) defined by Rhodes et al. (2017): the type dimension. In previous studies, this dimension has been analysed in terms of the type of sport practised (García et al. 2016), but in this chapter a quantitative element (the number of sports) is taken into account. This dependent variable has the features of a count variable (integer, non-negative and with a narrow range of variation). In the empirical application, the basic versions of count data models used in the microeconometric literature (Poisson and Negative Binomial models) are estimated, as well as the different specifications proposed for dealing with the excess of zeros problem, which is a controversial topic in the sports economics literature. The empirical results are discussed emphasizing the differences between the models, the interpretation of the coefficients and the relevance of making use of the calculation of the marginal effects in these non-linear models. The database used in the empirical analysis is the 2020 *Spanish Survey of Sporting Habits*, which is the most important source of information about the sports habits of the Spanish population.

The rest of the chapter is organized as follows. In the next section the different types of count data models are presented, paying special attention to the features of overdispersion and excess of zeros. Section 3 presents an overview of the economic literature on sports participation with a specific reference to the analysis of diversification. In the fourth section the data set and the definition of the variables used in the empirical application are described. In section 5 the empirical results of the estimation of different models of the number of sports practised are reported, discussing both the selection of the most appropriate model and the interpretation of the effects of the different explanatory variables. Section 6 concludes.

2. Econometric analysis of count data models

As mentioned above, the specific features of a count variable can be summarized by indicating that it is a nonnegative discrete (integer) variable and it usually takes a small number of low values including zero. The simplest distribution which accounts for these specific features is the Poisson distribution, whose probability mass function is given by:

$$Pr(y_i = Y_i) = P_{Y_i} = \frac{e^{-\mu_i}\mu_i^{Y_i}}{Y_i!} \qquad \mu_i > 0 \tag{1}$$

where the subindex *i* refers to the *ith* individual and μ is the only parameter associated with this distribution and it represents both the expected value and the variance. Consequently, the only way to introduce explanatory variables (*X*) in this count data model is to make μ dependent on these variables. The usual parameterization is:

$$\mu_i = exp(X_i^{'}\beta) \tag{2}$$

This model is estimated by maximum likelihood and the log of the likelihood function (*ln L*) in the case of using a random sample (**N** observations) is:

$$ln\, L(\beta) = \sum_{i=1}^{N}\left[Y_i\; X_i^{'}\beta - exp(X_i^{'}\beta) - ln(Y_i!)\right] \tag{3}$$

As happens when using these parametric approximations, the Poisson distribution imposes a specific pattern for the corresponding probability mass function. In particular, the ratio between the probabilities of occurrence of two consecutive values is given by:

$$\frac{P_{Y_i}}{P_{Y_i-1}} = \frac{\mu_i}{Y_i} \tag{4}$$

Notice that the value of μ determines the mode of the distribution. For instance, for μ between 0 and 1 the pattern is decreasing and the mode is zero. In general, for μ between *K* and *K+1*, the mode will be *K* and the pattern of the probabilities is first increasing with *y*, up to *K*, and then tending to zero at a fast rate.

The most relevant feature of the Poisson distribution is the equality of the expected value and the variance, known as the equidispersion property. In principle, this is a very restrictive property from an econometric point of view and it is not usually representative of the features of the count variables in most of the empirical applications. One way of dealing with the limitation of the Poisson model of not being appropriate for situations where there is overdispersion (i.e. the variance is greater than the expected value) in the count variable, a common feature of count data variables in economic models[2], is to look for a more flexible distribution in terms of the number of parameters that characterize it. The most used candidate is the Negative Binomial distribution, which can be defined straightforwardly or which can be motivated as a kind of extension of the Poisson distribution, highlighting the connection between the two models.

Suppose the parameter μ (now called μ* to use the same notations as in the econometric computer package *Stata*) in the Poisson model is now dependent on observed heterogeneity (X) but also on unobserved heterogeneity by means of a random variable (ν) such that:

$$\mu_i^* = \mu_i \nu_i = exp(X_i^{'}\beta)\nu_i \qquad \nu_i > 0 \tag{5}$$

Assuming that ν has a Gamma distribution[3] with parameters $1/\gamma$ and γ, which means that E(ν)=1 and var(ν)=γ, the marginal density of y is obtained by integrating out the random variable ν:

$$h(y_i \mid \mu_i, \gamma) = \frac{\Gamma(\gamma^{-1} + y_i)}{\Gamma(\gamma^{-1})\Gamma(y_i + 1)} \left(\frac{\gamma^{-1}}{\gamma^{-1} + \mu_i} \right)^{\gamma^{-1}} \left(\frac{\mu_i}{\mu_i + \gamma^{-1}} \right)^{y_i} \tag{6}$$

which corresponds to the Negative Binomial distribution, whose expected value and variance are:

$$\begin{aligned} &E(y_i) = \mu_i && \mu_i > 0 \\ &var(y_i) = \mu_i\left[1 + \gamma\,\mu_i\right] && \gamma > 0 \end{aligned} \tag{7}$$

Notice that the equidispersion property of the Poisson model does not hold insofar as γ is different from zero and the Poisson model is a particular case of the Negative Binomial model ($\gamma = 0$).

Depending on whether the parameter γ is assumed to be constant (α using *Stata* notation) or the variance is assumed to be proportional to the expected value, i.e. $\gamma = \delta/\mu_i$, which means $var(y_i) = \mu_i\,(1+\delta)$, two alternative versions of the Negative Binomial model

[2] Underdispersion is not as usual, although some examples can be found in the empirical literature, e.g. Lord et al. (2010). Hilbe (2014) presents the generalized Poisson model, which is an extension of the basic version that includes a dispersion or scale parameter allowing for underdispersion.

[3] Some other distributions, like the log-normal or the inverse-Gaussian distribution, can be used as an alternative to the Gamma distribution, but simulated maximum likelihood methods are required.

are considered in the empirical literature: type 2 (NB2) and type 1 (NB1), respectively. The most used model is the NB2, also known as the mean dispersion model, whereas the NB1 model is also referred to as the constant dispersion model. Testing for equidispersion (the Poisson model) reduces to testing whether α or δ are equal to zero depending on the type of Negative Binomial model you are using. This is done in *Stata* using a likelihood ratio test.

This model is estimated by maximum likelihood and the log likelihood function of the NB2 model in the case of using a random sample is:

$$ln\, L(\beta,\gamma)=\sum_{i=1}^{N}\left[ln\left(\frac{\Gamma(\gamma^{-1}+Y_i)}{\Gamma(\gamma^{-1})\Gamma(Y_i+1)}\right)+\gamma^{-1}\,ln\left(\frac{\gamma^{-1}}{\gamma^{-1}+exp(X_i^{'}\beta)}\right)+Y_i\,ln\left(\frac{exp(X_i^{'}\beta)}{exp(X_i^{'}\beta)+\gamma^{-1}}\right)\right] \tag{8}$$

These Poisson and Negative Binomial models are the count data counterparts of the Tobit model in the context of the limited dependent variable models. They deal with modelling a non-negative variable that takes the value zero for a significant proportion of observations by using the same model to explain whether the dependent variable is zero or not, and the positive values. They only differ in the kind of distribution of the dependent variable because of the integer values associated with a count variable.

In fact, one of the empirical issues associated with count variables is that the proportion of observations with a zero value for the dependent variable is greater than that which can be modelled assuming a Poisson or a Negative Binomial distribution, according to the structure of the probabilities implied by the corresponding probability mass functions. This is usually known in the literature as an excess of zeros problem. The usual way of dealing with this problem in the count data literature is to use the zero-inflated versions of the standard Poisson and Negative Binomial models. These are models that allow for the possibility of the zeros being generated by two processes: one generating zeros corresponding to potential non-participants and the second one allowing for some zeros being associated with optimal solutions (corner solutions in the Tobit terminology). Consequently, not only the excess of zeros, but also the limitation of the standard models, which use just one model to deal with the explanation of the zeros and the positive values of *y*, are taken into account.

The zero-inflated models include two equations. The first one is a binary discrete choice model fitting the probability of not being a potential participant (D_i=1), in which the dependent variable (D) is not observed, usually called the *inflate* equation. The second equation fits the number of times a particular event occurs for those who are potential participants, including the possibility of zero occurrences, by using a standard count data model. The two equations are assumed to be independent. This specification replicates the double-hurdle model with independent errors used in the limited dependent variable literature.

Consequently, the probability of observing a zero is equal to the probability of not being a potential participant (P_i) plus the probability of observing a zero (*F(0)*) when being a potential participant:

$$Pr(y_i = 0) = P_i + (1 - P_i)F(0) \qquad (9)$$

where the probability of not being a potential participant P_i is:

$$P_i = Pr(D_i = 1) = 1 - F^*(-Z_i'\delta) = F^*(Z_i'\delta) \qquad (10)$$

where F^* is the cumulative distribution function of a logistic (Logit model) or a standardized normal distribution (Probit model), Z is the vector of explanatory variables of the potential non-participation equation and δ is the vector of parameters. Notice that D is defined in the same way as in *Stata* in order to facilitate the interpretation of the results of the *inflate* equation.

The probability of observing a positive value of y is given by the probability of being a potential participant times the probability of y taking the positive value which is observed (Y_i), according to the probability mass function F of the corresponding standard model (Poisson or Negative Binomial) of the second equation, i.e.:

$$Pr(y_i = Y_i) = (1 - P_i)F(Y_i) \qquad if\ Y_i > 0 \qquad (11)$$

where the explanatory variables (X) of the second equation are introduced in the standard count data model through the parameter μ.

Consequently, the log of the likelihood function for these zero-inflated models is:

$$ln\, L(\beta,\delta,\gamma) = \sum_{Y_i=0} ln\big[P_i + (1 - P_i)\, F(0)\big] + \sum_{Y_i>0} \big[ln\,(1 - P_i) + ln\, F(Y_i)\big] \qquad (12)$$

where $\gamma=0$ in the zero-inflated Poisson model.

Notice that each of the standard count data models is a particular case of the corresponding zero-inflated version, when the probability of being a potential participant is one for all the individuals. The comparison between the standard and the zero-inflated versions can be made using any of the usual information criteria (Akaike, Bayesian) but, according to Wilson (2015), the use of the Vuong's test for non-nested models is not appropriate in this case.

In terms of how the zeros are generated, an alternative specification is the so-called hurdle model, which considers that all of them can be associated with being a non-potential participant, but not as an optimal zero. This model can be interpreted as the count data version of the two-part model. Two equations are considered for this particular model. The first one corresponds to a binary discrete choice model for explaining whether the count variable is positive or zero, which is observed. The second equation corresponds to a count data model where the distribution of the count variable is truncated at zero because it only applies to individuals for whom the number of times of occurrence of the particular event is positive.[4]

Consequently, the probability of observing a zero, i.e. not being a participant, is given by:

$$Pr(y_i = 0) = Pr(D_i^* = 0) = 1 - Pr(D_i^* = 1) = 1 - P_i = F^*(-Z_i^{'}\delta) \qquad (13)$$

where F^* is the cumulative distribution function of a logistic (Logit model) or a standardized normal distribution (Probit model), D^* is a dummy variable equal to 1 if the individual participates (y_i >0) and Z is the vector of explanatory variables of the participation equation.

And the probability of observing a certain positive value of y is:

$$Pr(y_i = Y_i) = P_i \frac{F(Y_i)}{1 - F(0)} \qquad if\ Y_i > 0 \qquad (14)$$

where the standard distribution F is truncated at zero and the explanatory variables (X) of the second equation are introduced in the standard count data model through the parameter μ.

[4] See Feng (2021) and Green (2021) for a comparison of zero-inflated and hurdle models.

The log likelihood function of this hurdle version is:

$$ln\, L(\beta,\delta,\gamma) = \sum_{Y_i=0} ln\,(1-P_i) + \sum_{Y_i>0} \left[ln\,(P_i) + ln\, F(Y_i) - ln\left(1-F(0)\right)\right] \qquad (15)$$

where γ=*0* in the hurdle Poisson model.

Notice that this likelihood function can be maximized by estimating separately the discrete choice model for *D** and the truncated count data model for the observations with a positive value of the dependent variable (*y*). Additionally, it is worth mentioning that the hurdle model is not a particular case of the corresponding zero-inflated version because a truncated model, and not a standard version, is estimated for those observations with a positive value for the number of times of occurrence of a particular event.

As an alternative to the usual count data models, ordered models (McCullagh 1980) could also be considered. Although they are intended for situations in which the dependent variable is a qualitative variable whose categories can be ordered according to the characteristic which is measured, they could be used for count variables where, to some extent, the interest is in explaining the probability of occurrence of a particular event and the values of the count variable can be considered as ordered.

The basic idea behind this model is that there is a latent variable (*y**), which is not observed by the econometrician, which measures the intensity of occurrence of a particular event. What is observed is a discrete indicator, which approximates the level of this intensity, i.e. the number of times a particular event occurs - the observed (count) variable.

The structure of the ordered model is as follows:

$$\begin{aligned} &y_i^* = X_i^{'}\beta + \varepsilon_i \\ &y_i = j \qquad if\;\; \kappa_j < y_i^* \le \kappa_{j+1} \quad j = 0,....,J \\ &\qquad\qquad \kappa_0 = -\infty \quad \kappa_{J+1} = \infty \end{aligned} \qquad (16)$$

where ε_i is the error term, whose distribution determines whether a Logit (logistic) or a Probit (standardized normal) is considered and κ_j are the cut points defining the intervals associated with the different values of the dependent variable (*y*) and are parameters to be estimated.

The probability of *y* taking a particular value *j* is:

$$Pr(\, y_i \,=\, j\,) = Pr\left(\kappa_j < y_i^* \leq \kappa_{j+1}\right) = F\left(\kappa_{j+1} - X_i^{'}\beta\right) - F\left(\kappa_j - X_i^{'}\beta\right) \qquad (17)$$

where *F* is the cumulative distribution function corresponding to the error term. This is the contribution to the likelihood function of each observation in the sample.

The ordered model has the advantage of being more flexible in fitting these probabilities than the models mentioned previously, since there is no specific pattern for them, as happens in the case of the standard count data models. As with the previous specifications, the ordered model can be estimated through a single equation, treating zero as an additional category, but there is also a hurdle version, which separates the decisions to participate and the number of sports to perform, and a zero-inflated version as in the case of count data models (Harris and Zhao 2007).

3. Sports participation and its diversification

There is plenty of scientific evidence about the beneficial effects of regular physical activity on health (Rhodes et al. 2017). However, sedentary lifestyles have become a major problem in developed countries, to the extent that the World Health Organization has developed the *Global Action Plan on Physical Activity 2018-2030*, which sets the goal of reducing physical inactivity by 10% by 2025 and 15% by 2030. This plan connects with the 2030 Sustainable Development Goals (SDGs), an initiative promoted by the United Nations to boost economic growth taking into account environmental, economic and social sustainability.

There is an extensive economics literature on the determinants of individual sports practice. A survey of the main theoretical approaches, the correlates of sports participation and its effect on health, well-being, social capital and labour market outcomes has been done by Cabane and Lechner (2015). The literature review reveals that there is considerable heterogeneity in the definition of the variable under study. In fact, as Rhodes et al. (2017) indicate, physical activity has multiple dimensions: frequency, intensity, time and type (FITT principles). Most economic studies about the correlates of physical activity analyse participation, time and/or frequency of practice. The intensity of practice has been less studied in the literature and, regarding type, many researchers add any type of sport or physical activity whereas others examine specific sports, groups of sports with common characteristics, or physical activity done in different domains of daily life (leisure, work, commuting, etc.).

One of the characteristics of the variables that measure sports practice is that a high percentage of people do not play sports, so it becomes necessary to model the zeros in the dependent variable. The lack of participation may be due to different reasons: preferences, time or monetary constraints, or infrequency of practice, i.e. when the period of time considered in the survey is not long enough to capture sports practice. The reasons for observing zeros may determine the econometric specification.

This chapter focuses on the type dimension of sports, measured by the number of sports the individual has practised in the previous year. The correlates of individual diversification of the sports activity are studied by estimating count data models to take into account the specific features of the dependent variable (integer, non-negative and with a narrow range of variation).

The individual diversification of sports activity -also called multipractice or omnivorousness- has hardly been analysed in the sports economics literature, but it has received more attention in the sociology and sport science fields. In some cases, diversification is studied to contribute to the debate about which is the best way to attain elite sport performance: either to specialize in one sport at early ages or to sample different sports as a child and specialize a little later (Bridge and Toms 2013). Other researchers analyse differences in diversification between countries (Lefèvre et al. 2021) or between social classes (Lefèvre and Ohl 2012). Although globalization could lead to the homogenization of tastes and behaviours, diversification in sports may vary between different countries with a different socio-cultural environment or it may be a way to differentiate oneself from others in lower social positions. In this regard, the question is whether higher social classes have a more omnivorous behaviour in their sporting activities than lower classes have, as some authors have found in the study of cultural demand. In particular, Lefèvre and Ohl (2012), using French data, obtain that, although it is frequent for people to practise more than one sport, higher social groups tend to be more omnivorous than the rest. Regarding differences between countries, on the one hand, Lefèvre et al. (2021) find that Japanese are less likely to do sports than the French and those who participate are more likely to be single-sport practitioners. On the other hand, Lefèvre et al. (2020) compare France and Spain, and conclude that there are similarities in the factors associated with the number of sports practised: males, younger people, more educated people and those with sporty fathers tend to have a greater portfolio size.

Regarding the econometric methodology, count data models have already been applied in the literature on sports participation.[5] This is the case of Dawson and Downward (2011), who estimate different count data specifications for the UK (Poisson, Negative Binomial –NB-, Zero-inflated Poisson –ZIP- and Zero-Inflated Negative Binomial –ZINB-) to study the number of hours and the number of days of sports participation. According to their results, the ZINB specification is the most suitable model for their data. Muñiz et al. (2014) and García et al. (2016) also estimate ZINB models to analyse the number of times that Spanish individuals play different types of sports. Anokye et al. (2012) opt for the Negative Binomial to model the number of days of sports practice in the previous four weeks by participants at a British university. Oliveira-Brochado et al. (2017) specify a two-part model –specifically a Negative Binomial hurdle model- to study participation and frequency (number of days per week) of sport by Portuguese people and, more recently, García and Suárez (2021, 2023) also estimated two-part models for the number of days per week allocated to sports by Mexicans, where the frequency of practice by participants is modelled through truncated Poisson and Negative Binomial models.[6] Previous papers, by Slymen et al. (2006) and Lee at al. (2016), used zero-inflated models for the number of days per week practising physical activity by Hispanic females in San Diego (USA) and by patients with Parkinson's disease in South Korea, respectively. In a similar context of analysis of diversification, Lefèvre et al. (2020) apply a Negative Binomial hurdle model to analyse the number of sports practised by French and Spanish individuals.

4. Data

The empirical analysis uses the most recent edition of the *Spanish Survey of Sporting Habits*, corresponding to 2020. This survey is conducted every five years and the last

[5] See Downward and Muñiz (2019) for a brief review of the econometric approaches used in the analysis of sports participation, where count data models are included.

[6] In García and Suárez (2020) a model for the number of days per week of sports practice is estimated using a version of the ordered model where the cut-points are known and a lognormal distribution is specified because their model is a kind of two-part model.

edition has been carried out by the Spanish Ministry of Culture and Sports together with the Higher Sports Council. The information was collected between September 2020 and January 2021. At that time, the lockdown due to the COVID-19 pandemic had already ended a few months earlier.

The survey is aimed at people aged 15 or over and gathers information about sporting habits of the Spanish population, not only about their sports practice but also about attendance at sports events, sports equipment available at home, interest in sports information, relationship with sports organizations, etc.

According to the data provided by the survey, around 60% of Spaniards have played sports in the previous year and, among those who did sports, the vast majority (92%) practised it at least once a week. Regarding the number of sports done by participants (variable *ndep* in the *Stata* commands reported in the next section), and focusing on the sample selected for the empirical analysis (where those respondents who are incapacitated for work or who state that they have practised more than 30 sports in the previous year are dropped),[7] Figure 1 shows that most of them do one or two activities. In fact, it is more frequent to practise two activities (26% of participants) than one (24%). The same occurs in France but not in Japan, where single-practitioners predominate (Lefèvre et al. 2021). Moreover, less than 15% of the individuals who play sports do more than five activities. Thus, the distribution of the dependent variable is right-skewed and it is concentrated in a small range of values.

<< INSERT FIGURE 1 HERE >>

The most practised sports by participants are, in order of importance: light gymnastics, cycling, intense gymnastics, mountaineering or hiking, athletics (running, walking, etc.) and bodybuilding. All these sports, except bodybuilding, also occupy the first positions among single-practitioners, although the ranking changes somewhat. However, the classification for those who play two sports is almost identical to the general ranking.

As for the most frequent combinations of sports, and focusing on those who practise two activities, the most common pattern is to combine light gymnastics with

[7] Firstly, people who are incapacitated for work may not be able to play sports. Secondly, it is quite implausible to have played more than 30 sports. These restrictions reduce the sample size by around 1.6%.

mountaineering or hiking, intense gymnastics, athletics, cycling or other sports not explicitly mentioned in the questionnaire.

The explanatory variables used in the empirical exercise of the next section are as follows (in square brackets the name of these variables in the *Stata* commands reported in the next section):

- *Gender*: male, female [*sexo*]
- *Age*: five age groups (younger than 25, 25-34, 35-44, 45-54, older than 54) [*g_edad*]
- *Education*: less than primary school, primary school, secondary school level 1, secondary school level 2, university degree [*estud_new*]
- *Employment status*: worker, unemployed, retired, student, housework, other situation [*sitlab*]
- *Marital status*: single, married, other situation [*sitpers_new*]
- *Health status*: very good, good, fair, bad, very bad [*estfis*]
- *Parents practising sport*: yes, no [*pract_padres*]
- *Nationality*: Spanish, double nationality, foreigner [*nac*]
- *Municipality*: provincial capital, more than 50,000 inhabitants, fewer than 50,000 inhabitants [*tram_munic*]

Table 1 shows some descriptive statistics of these variables. Males, young people and persons with secondary or university education are more likely to do sports. The same can be said of students, single persons, people in very good health and those whose parents have played sports. Finally, it is interesting to note that the diversity of practice differs between sporty people depending on their sociodemographic characteristics. This information is included in Table 1. The figures show that the average number of activities is higher among males, young people, individuals with higher education level, students, single, and people in very good health.

<< INSERT TABLE 1 HERE >>

5. Results

The estimates of the different models discussed in the methodology section, applied to the number of sports practised by people in Spain, are reported in Tables 3 and 4. First, the fit and the appropriateness of the different specifications are discussed, and then the comments on the interpretation of the coefficients and the calculation of the marginal effects of the explanatory variables on the dependent variable follow.

The empirical analysis has been done using *Stata* software. Table 2 shows the commands and the syntax used to estimate each of the models. More details about how to use *Stata* to estimate and analyse the different count data models discussed in this chapter can be found in Cameron and Trivedi (2022). The estimation of these models is also possible using other major statistical software packages. Friendly and Meyer (2015) provide procedures to estimate different count data models in *R*, whereas Hilbe (2014) shows the codes for running these models in *R*, *SAS* and *Stata*. Other software packages such as *SPSS* or *LIMDEP* include procedures to estimate count data models too.

<< INSERT TABLE 2 HERE >>

<< INSERT TABLE 3 HERE >>

<< INSERT TABLE 4 HERE >>

5.1. Comparison of the different models

The descriptive analysis of the dependent variable provides relevant information about the appropriateness of the standard models. As indicated in Figure 1, the sample mean of the number of sports (1.92) is clearly smaller than the sample variance (6.76), which means that this variable has overdispersion and, consequently, the fit of the standard Poisson model should be worse than that of the Negative Binomial model.[8] In fact, the values of the log likelihood function reported in Table 3 confirm that the fit of the Negative Binomial Type 2 model (henceforth Negative Binomial) is much better. This evidence is corroborated by the fact that the estimate of the overdispersion parameter (α) is significantly different from zero, rejecting the Poisson specification.

In order to highlight the differences between models, it is useful to look at the adjusted probabilities for each value of the number of sports variable across models. The mean values of these adjusted probabilities and the corresponding relative frequencies in the sample are reported in Table 5. It is evident that the standard Poisson model is not able to reproduce the proportion of observations corresponding to individuals not

[8] Results for the Negative Binomial Type 1 model are not reported here since it is hardly used in the empirical economic literature and the results point in the same direction as those of the Type 2 version.

practising sports (22.58% versus 38.85% in the sample), whereas the Negative Binomial model captures this proportion better. In the case of the Poisson model, the average probability of practising one sport is higher than that of not practising sports, and then it follows a decreasing pattern as the number increases. This is an expected result since the sample mean of the number of sports, which is an estimate of μ – the parameter of the Poisson distribution - is below 2 and, according to the result in expression (3), the mode should be 1, as happens in the adjusted probabilities of the Poisson model. Notice also that the Negative Binomial model is not able to replicate the observed pattern of the relative frequency (higher for y=2 than for y=1 and y=3) because of the particular features of this distribution.

<< INSERT TABLE 5 HERE >>

The zero-inflated versions of the standard models are designed to capture the potential problem of excess of zeros and/or the fact that the reason behind observing a zero could be either being a non-potential participant or an optimal choice by a potential participant. In Table 4, the estimated zero-inflated versions of the Poisson and the Negative Binomial models are reported. There is a substantial improvement in the value of the log likelihood model in the case of the zero-inflated Poisson model when compared to the standard version, as a consequence of the bad performance of the latter in estimating the probability of not practising sport. The zero-inflated version of the Negative Binomial model also has a higher value of the log likelihood compared to its standard version, and it is also preferred to the zero-inflated Poisson version because the overdispersion parameter α is significantly different from zero.

As mentioned in the methodology section, strictly speaking, the Vuong's test cannot be used to check the appropriateness of the standard versions versus the zero-inflated versions, but the change in the log likelihood values for both the Poisson and the Negative Binomial versions is so important that the zero-inflated versions are clearly preferred in terms of any of the usual information criteria. In fact, as it can be seen in Table 4, the second equation of the zero-inflated models (number of sports) does not include health status as an explanatory factor. This is because, when using the same set of variables in both equations, there were some problems with the estimation of the parameters, and their standard errors, of the health status variables and this kind of exclusion restriction was used to obtain convergence.

The adjusted probabilities reported in Table 5 show that the zero-inflated versions are able to replicate the proportion of individuals who do not play sports. Moreover, the zero-inflated Poisson model reproduces the pattern of the relative frequencies associated with practising one, two and three sports better than the Negative Binomial model does, but it clearly underestimates the frequencies for one and two, and overestimates those corresponding to values between three and six, and then the estimated probabilities tend to zero more rapidly than the observed frequencies. In the case of the zero-inflated Negative Binomial model, the adjusted probabilities decrease with the number of sports practised from the start, not capturing the pattern of the relative frequencies in the sample for values one, two and three, but fitting quite well frequencies above three and the proportion of individuals practising one, two or three sports (42.52% in the sample and 40.38% estimated by the Negative Binomial model, compared to the 37.83% estimated by the Poisson model).

As discussed in the methodology section, an alternative specification is the hurdle count data model, a two-part model in which all zeros come from potential non-participants and the second equation (number of sports) is estimated by using a truncated (at zero) distribution for the count variable. The estimates of the hurdle versions of the Poisson and the Negative Binomial model are reported in Table 4 using the same specification as for the zero-inflated versions, i.e. excluding health status in the second equation. Comparing the results shown in Table 4, the estimated coefficients differ between the hurdle and the zero-inflated models. This is because the first hurdle has a different definition depending on whether the zero-inflated or the hurdle version is considered, i.e. depending on how the zeros are generated. But the overall value of the log likelihood function is almost exactly the same for both versions of either the Poisson (-9493) or the Negative Binomial model (-9020), although it is marginally better for the hurdle models. This happens for this particular empirical exercise, but it cannot be generalized. In fact, when simplifying the specification by reducing the explanatory variables of the second equation, the difference between the two log likelihood values is more significant in favour of the hurdle model.

The main point about the above result is that, although the two models seem to have similar explanatory power, they provide a different interpretation about how the zeros are generated and, consequently, the decision about which model to choose has to be based on whether the characteristics of the data set (the period of reference associated with the question about the number of sports practised) can be related to one or two potential sources of zeros. In this particular case, the question refers to the number of sports practised during a year and zeros should be better associated with potential non-participants than with optimal choices made by people who are willing to practise sports. Notice that the similar performance of the zero-inflated and the hurdle models translates into almost exactly the same estimates for the probabilities of practising a given number of sports, according to the values reported in Table 5.

It is also relevant to remember that the hurdle model is not nested into the corresponding zero-inflated version. This is because the second equation requires a distribution of a positive random variable for the dependent variable, the truncated (at zero) Poisson or Negative Binomial, whereas being nested would require assuming the unrestricted standard versions of these distributions. In the case in which this “nested” model is estimated, the log likelihood of the hurdle Negative Binomial model is -9442, smaller than that of the zero-inflated model.

Finally, as mentioned above, the estimation of a count data model can be performed by using ordered discrete choice models. They have the advantage over the previous count data models that they do not impose any restriction on the adjusted probabilities associated with the values of the dependent variable. As is reported in Table 3, the standard ordered Logit model has a significantly better fit than the standard Poisson and Negative Binomial models. This is also the case when considering the estimates of the hurdle version of the ordered model, whose estimates are shown in the last column of Table 4. It should be mentioned that the dependent variable has been redefined in the case of those observations for which the number of sports is greater than 20 (four of them) (variable *ndep_o* in the Stata commands for the ordered models). The last category of the dependent variable in the ordered models corresponds to the case in which the number of sports is equal to or greater than 20. If this redefinition is not applied, the log likelihood increases to -8958, still greater than that of the count data models. This flexibility of the ordered model becomes evident when looking at the adjusted probabilities in Table 5. In the case of the ordered hurdle model, they reproduce almost exactly the relative frequencies in the sample and, in particular, the specific profile of the probabilities associated with practising one, two and three sports.

5.2. Interpretation of the coefficients and the marginal effects

Table 3 reports the coefficients of the explanatory variables for the standard models, corresponding to the elements of vector β in (2) or (5). According to the parameterization for the count data models, a positive coefficient implies a positive effect of the corresponding variable on the expected value of the number of sports, as in the case of the ordered model in terms of the latent variable y^* in (16).

In this regard, females have a lower expected value of the number of practised sports compared to males because of the negative sign of the coefficient of this dummy. In the case of age, an interpretation of each individual coefficient can be made, concluding, for instance, that those who are in the age group 35-44 play a lower expected number of sports than those younger than 25 (the reference category). But also an overall interpretation of the pattern of the age coefficients can be made, since the categories are ordered. Thus, age has a negative effect on the expected number of sports, not because the coefficients are negative but because the pattern they follow is decreasing. This also applies to the health status variable, whose categories are also ordered: the worse the health status the lower the expected value of the number of sports. Regarding other variables, education has a positive effect on the expected value of the dependent variable, as happens for those who live in small municipalities and those whose parents practise sport. Those who are not single are expected to practise fewer sports, as it is the case of those who are not Spanish. Finally, people who are either unemployed, retired or doing housework have a lower expected number of sports compared to those who are working –the reference category-, whereas those who are students or are in the default category have a higher expected value.

We can also make an interpretation of the size of the effect of a unit change in each explanatory variable because, given expressions (2) and (5) for the count data models, increasing the *k-th* explanatory variable by one unit will multiply the expected number of sports by exp(β_k), which, for "small" values of β_k, means that the expected number of sports will increase by (β_k x100)%.[9] This means that in the Poisson model females have an expected value of the number of sports exp(-0.277) times that of males with the same characteristics (or exp(-0.294) for the Negative Binomial model), i.e. the expected value

[9] Notice that by taking logs in (2) and (5), we end up with an expression similar to the log-linear version of the regression model and the interpretation is the same as mentioned above.

is 24.2% lower for females than for males. In the case of age, the coefficient for the group 25-34 implies that the expected value for the number of sports is 8.8% lower than that of people younger than 25 (this is a “good” approximation insofar as exp(-0.088) is 0.916).

Notice that the marginal effects, interpreted as the absolute change in the expected value of the dependent variable, are different for each individual because of the non-linearity of expressions (2) or (5), as becomes evident from the following expression, which corresponds to the effect of a continuous variable.[10]

$$\frac{\partial E(y_i | X_i)}{\partial X_i^k} = \beta_k \exp(X_i^{'}\beta) \qquad (18)$$

In the case of qualitative variables, such as all those used in this exercise, the marginal effects are calculated as the change in the expected value of *y* when the dummy defined for each specific category changes from 0 to 1, i.e. when there is a change from the reference category to that associated with the dummy. In the first rows of Table 6 the average of the marginal effects of gender and health status on the expected value of the number of sports for all individuals in the sample are reported.[11] They look relatively similar for both the Poisson and the Negative Binomial model, despite having different explanatory power, but the difference is more significant in the case of fair health status because of the greater difference in the size of the coefficients in the two models. In the case of the ordered Logit model, the marginal effects are calculated by using the adjusted value of the expected number of sports obtained as the sum of the products of the adjusted probabilities times the different values of the number of sports.

<< INSERT TABLE 6 HERE >>

But, as mentioned in the methodology section, the analysis of the effects of the covariates on the probabilities associated with the different values of *y* is as important as, or even more important, than studying the marginal effects on the expected value of *y*. In fact, this is the aspect where differences between models are more evident. Table 6 offers information on the marginal effects of gender and health status on the probabilities of practising zero to four sports. Differences between the Poisson and the Negative Binomial

[10] There are several ways of reporting marginal effects. They can also be expressed as elasticities or semi-elasticities depending on the features of the explanatory variable. *Stata* allows for the possibility of choosing how the marginal effects are calculated.

[11] In some papers marginal effects are evaluated at the mean values of the explanatory variables. Consequently, when writing up the results of empirical research, it is important to make clear how the marginal effects are calculated.

model can be clearly appreciated, as well as differences with respect to the ordered Logit model.

Turning now to the zero-inflated and hurdle models, whose estimates are reported in Table 4, it is worth remembering that both are two-equation models but they differ in the role of the first equation: to explain whether or not the individual is a potential sports practitioner (zero-inflated model) or just whether the individual practises sport or not (hurdle model). This feature has to be taken into account when interpreting the results of the Logit models corresponding to this first stage. Additionally, the estimates of the first equation in the double-hurdle model (*Inflate*) correspond to defining the depending variable of the binary discrete choice model as equal to 1 if the individual is a potential non-participant and zero otherwise, as is done in the output produced by *Stata*. Regarding the interpretation of the estimated coefficients in the second equation, it must be done in terms of the effect of the corresponding variable on the expected number of sports for potential practitioners in the zero-inflated models, and conditional on being a practitioner in the hurdle model, proceeding as it was done when discussing the effects for the standard models.

Although this is not necessarily the case in all the empirical applications, in this particular exercise the coefficients of each explanatory variable have the same sign in both equations of all the versions of the zero-inflated and hurdle models, i.e. a variable which has a positive effect (positive coefficient) on the expected number of sports also has a positive effect on the probability of being a potential practitioner in the zero-inflated model (negative coefficient of the *inflate* equation) and on the probability of practising sport in the hurdle model. But the significance of those effects differs. For instance, the coefficients capturing the effect of the individual employment status are in most cases significant in the second equation (number of sports) of the zero-inflated model but not in the *inflate* equation, which is also the case to some extent for the hurdle equation. By contrast, the effect of education on the unconditional expectation of the number of sports works mainly through the probability of being a sports practitioner in the different versions of the hurdle model. The existence of different patterns depending on the equation considered is relevant in terms of the policies to be implemented to promote sports participation, depending on whether the aim is just to increase the number of sports practitioners or the diversification of this participation, measured in this example by the number of sports practised. Notice that this differentiated pattern cannot be captured by the standard versions of the different models.

Table 6 reports the average of the marginal effects of gender and health status for these two-equation models on the unconditional probabilities and the unconditional expected number of sports for all the individuals in the sample. In the case of the zero-inflated count data models the unconditional expectation of *y* is:

$$E(y_i|\, X_i, Z_i) = exp(X_i^{'}\beta)\, Pr(D_i = 0) = exp(X_i^{'}\beta)\,(1-P_i) = exp(X_i^{'}\beta)\, F^*(-Z_i^{'}\delta) \qquad (19)$$

where P_i is defined as in (10).

Similar to expression (18), the marginal effects on the unconditional expectation of the dependent variable for the zero-inflated count data models of the vectors X_i and Z_i are given by the following expressions:

$$\frac{\partial E(y_i|\, X_i, Z_i)}{\partial X_i^k} = \beta_k\; exp(X_i^{'}\beta)\, F^*(-Z_i^{'}\delta) \qquad (20)$$

$$\frac{\partial E(y_i|\, X_i, Z_i)}{\partial Z_i^k} = -\,\delta_k\; exp(X_i^{'}\beta)\;\; \frac{\partial F^*(-Z_i^{'}\delta)}{\partial Z_i^k} \qquad (21)$$

where *F** is defined in (10). When a variable is included in both X_i and Z_i, its marginal effect is the sum of the expressions (20) and (21).

In the case of the hurdle count data models the unconditional expectation of *y* is:

$$E(y_i|\, X_i, Z_i) = exp(X_i^{'}\beta)\, Pr(D_i^* = 1) = exp(X_i^{'}\beta)\, P_i = exp(X_i^{'}\beta)\,(1\text{-}\, F^*(-Z_i^{'}\delta)) \qquad (22)$$

where P_i is defined as in (13).

The marginal effects on the unconditional expectation of the dependent variable for the hurdle count data models of the vectors X_i and Z_i are given by the following expressions:

$$\frac{\partial E(y_i|\, X_i, Z_i)}{\partial X_i^k} = \beta_k\; exp(X_i^{'}\beta)\,(1\text{-}\, F^*(-Z_i^{'}\delta)) \qquad (23)$$

$$\frac{\partial E(y_i|\, X_i, Z_i)}{\partial Z_i^k} = \delta_k\; exp(X_i^{'}\beta)\;\; \frac{\partial F^*(-Z_i^{'}\delta)}{\partial Z_i^k} \qquad (24)$$

Notice that the marginal effect of those variables which are included in both X_i and Z_i is the sum of the expressions (23) and (24).

It is worth pointing out that, although the marginal effects on the expected value of *y* are very similar across models, a substantial difference is observed between the standard and the two-equation versions when considering the marginal effect on the unconditional probabilities, in particular, on the probability of not being a sports practitioner (y = 0). For instance, having a fair health status increases this probability by approximately 18 percentage points compared to having a very good health status in the two-equation models, and by around 7 percentage points in the standard count data models. Focusing on the two-equation models, there are some differences between the Poisson versions and the rest when looking at the probabilities of practising one sport or more than two sports, as a consequence of the equidispersion restriction associated with this model.

Finally, notice that the hurdle version of the ordered model, which has the best fit because of its flexibility, produces marginal effects that are quite similar to those of the Negative Binomial version of the hurdle model, which is the preferred specification among the typical count data models. In addition, although the fit is almost the same, the marginal effects of the Negative Binomial version of the hurdle model differ with respect to those of the zero-inflated model, probably as a consequence of the different assumptions about how the zeros are generated.

6. Concluding remarks

In this chapter a review of the most commonly used count data models when using cross-section data is presented. The dependent variable is characterized as an integer non-negative variable, which takes a small number of (low) values, usually with a variance greater than the expected value (overdispersion) and with a relatively large proportion of zeros (excess of zeros). Particular attention is devoted to the specification of the different models used in the empirical literature in terms of how the zeros, which are observed, have been generated. In this regard, the zero-inflated versions of the standard models are discussed, and the ordered models are presented as an alternative way of estimating these models in which the dependent variable represents the number of times something happens (e.g. the number of visits to the doctor).

All the methodological issues are illustrated by means of an empirical application to the analysis of diversification in sports practice, i.e. how many different sports are practised by individuals. This specific feature of sports participation has received little attention in the sports economics literature.

Both the hurdle and the zero-inflated versions of the Negative Binomial model are performing much better than the standard Poisson and Negative Binomial models. In this particular exercise, taking into account the characteristic of the survey used in the empirical analysis, the hurdle model seems to be more adequate.

It is worth emphasizing the importance of the analysis of the marginal effects of the different explanatory variables in these models, given their non-linearity. This is an issue that usually does not receive too much attention in empirical work. In particular, when dealing with count data models, attention should be paid, not only to the marginal effects on the expected value of the dependent variable, but also to the marginal effects on the probabilities associated with the values the dependent variable can take. Most of the differences in the performance of the alternative models are more evident when analysing the adjusted probabilities, since the distributional and the specification assumptions have more impact on the structure of the adjusted probabilities than in the expected value.

Finally, it is also relevant to mention some recent developments related to new models which have proposed to deal with the overdispersion (e.g. Altun 2019; Cahoy et al. 2021; Sellers and Premeaux 2021) and new proposals for clustered count data (e.g. Altinisk 2022).

References

Agrawal M, Sensarma R (2007) Determinants of merger activity: Evidence from India. Int J Financ Serv Manag 2(4):277-288. doi:10.1504/IJFSM.2007.016285

Altinisik Y (2022) Addressing overdispersion and zero-inflation for clustered count data via new multilevel heterogeneous hurdle model. J App Stat 50(2):408-433.
doi: 10.1080/02664763.2022.2096875

Altun E (2019) A new model for over-dispersed count data: Poisson quasi-Lindley regression model. Math Sci 13(3):241-247. doi: 10.1007/s40096-019-0293-5

Anokye NK, Pokhrel S, Buxton M, Fox-Rushby J (2012) The demand for sports and exercise: results from an illustrative survey. Eur J Health Econ 13(3):277-287. doi:10.1007/s10198-011-0304-4

Bridge MW, Toms MR (2013) The specialising or sampling debate: a retrospective analysis of adolescent sports participation in the UK. J Sport Sci 31(1):87-96. doi:10.1080/02640414.2012.721560

Cabane C, Lechner M (2015) Physical activity of adults: A survey of correlates, determinants and effects. J Econ Stat 235(4-5):376-402. doi:10.1515/jbnst-2015-4-504

Cahoy D, Di Nardo E, Polito F (2021) Flexible models for overdispersed and underdispersed count data. Stat Pap 62(6):2969-2990. doi: 10.1007/s00362-021-01222-7

Cameron AC, Trivedi PK (1986) Econometric models based on count data: comparisons and applications of some estimators and tests. J Appl Econom 1(1):29-53. doi:10.1002/jae.3950010104

Cameron AC, Trivedi PK (2005) Microeconometrics: Methods and Applications. Cambridge University Press, New York. doi: 10.1017/CBO9780511811241

Cameron AC, Trivedi PK (2014) Regression analysis of count data. Cambridge University Press, New York. doi:10.1017/CBO9781139013567

Cameron AC, Trivedi PK (2022) Microeconometrics using Stata. Volume II: Nonlinear models and causal inference methods, 2nd edn. Stata Press, Texas.

Cragg JG (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39(5):829-844. doi:10.2307/1909582

Dawson P, Downward P (2011) Participation, spectatorship and media coverage in sport: some initial insights. In: Andreff W (ed) Contemporary issues in sports economics: participation and professional team sports. Edward Elgar, Cheltenham, p 15-42. doi:10.4337/9780857930385

Downward P, Muñiz C (2019) Sports participation. In: Downward P, Frick B, Humphreys BR, Pawlowski T, Ruseski JE, Soebbing BP (eds) The SAGE Handbook of Sports Economics. SAGE Publications, London, p 33-44

Dupuy JF (2018) Statistical methods for overdispersed count data. Elsevier, Oxford. doi:10.1016/C2017-0-00831-5

Feng CX (2021) A comparison of zero-inflated and hurdle models for modelling zero-inflated count data. J Stat Distrib Appl 8(1):1-19. doi:10.1186%2Fs40488-021-00121-4

Friendly M, Meyer D (2015) Discrete data analysis with R: visualization and modeling techniques for categorical and count data. CRC Press, Boca Raton.

García J, Muñiz C, Rodríguez P, Suárez MJ (2016) Comparative analysis of sports practice by types of activities. Int J Sport Financ 11(4):327-348

García J, Suárez MJ (2020) Organised and non-organised after-school physical activity among children in Spain: the role of school-related variables. Eur Sport Manag Q 20(2):171-188. doi:10.1080/16184742.2019.1594329

García J, Suárez MJ (2021) Dimensions of sports participation: evidence from Mexico. In: Koning RH, Késenne S (eds) A Modern Guide to Sports Economics. Edward Elgar, Cheltenham, p 226-239. doi:10.4337/9781789906530

García J, Suárez MJ (2023) The relevance of specification assumptions when analyzing the drivers of physical activity practice. Econ Model 119: 106127. doi:10.1016/j.econmod.2022.106127.

Geil P, Million A, Rotte R, Zimmermann KF (1997) Economic Incentives and Hospitalization in Germany. J Appl Econom 12:295-312. doi:10.1002/(SICI)1099-1255(199705)12:3%3C295::AID-JAE443%3E3.0.CO;2-X

Green JA (2021) Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and Negative Binomial regression. Health Psychol Behav Med 9(1):436-455. doi:10.1080/21642850.2021.1920416

Harris MN, Zhao X (2007) A zero-inflated ordered probit model, with an application to modelling tobacco consumption. J Econometrics 141:1073-1099. doi:10.1016/j.jeconom.2007.01.002

Hausman J, Hall BH, Griliches Z (1984) Econometric Models for Count Data with An Application to the Patents-R&D Relationship. Econometrica 52(4):909-938. doi:10.2307/1911191

Hilbe, JM (2014) Modeling count data. Cambridge University Press, Cambridge. doi:10.1007/978-3-642-04898-2_369

Inan T (2021) Using Poisson model for goal prediction in European football. J Hum Sport Exerc 16(4):942-955. doi:10.14198/jhse.2021.164.16

Jang TY (2005) Count Data Models for Trip Generation. J Transp Eng 131(6):444-450. doi:10.1061/(ASCE)0733-947X(2005)131:6(444)

Lee J, Park CG, Choi M (2016) Regular exercise and related factors in patients with Parkinson’s disease: Applying zero-inflated negative binomial modelling of exercise count data. Appl Nurs Res 30:164-169. doi:10.1016/j.apnr.2015.08.002

Lefèvre B, Nohara H, Nier O (2021) Sports practice in Japan and France: A comparative analysis. Plos One 16(6):e0253435. doi:10.1371%2Fjournal.pone.0253435

Lefèvre B, Ohl F (2012) Consuming sports: distinction, univorism and omnivorism. Sport Soc 15(1):44-63. doi:10.1080/03031853.2011.625276

Lefèvre B, Routier G, Llopis-Goig R (2020) Sports participation in France and Spain: An international comparison of voraciousness for sport. Poetics 81:101429. doi:10.1016/j.poetic.2019.101429

Lord, D, Geedipally, SR, Guilkema, SD (2010) Extension of the application of the Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion. Risk Anal 30(8):1268-1276. doi:10.1111/j.1539-6924.2010.01417.x

McCullagh P (1980) Regression Models for Ordinal Data. J Roy Stat Soc B 42(2):109-142. doi:10.1111/j.2517-6161.1980.tb01109.x

Muñiz C, Rodríguez P, Suárez MJ (2014) Sports and cultural habits by gender: an application using count-data models. Econ Model 36:288-297. doi:10.1016/j.econmod.2013.09.053

Oliveira-Brochado A, Quelhas Brito P, Oliveira-Brochado F (2017) Correlates of adults’ participation in sport and frequency of sport. Sci Sport 32(6):355-363. doi:10.1016/j.scispo.2017.03.005

Pohlmeier W, Ulrich V (1995) An Econometric Model of the Two-Part Decision making Process in the Demand for Health Care. J Hum Resour 30(2):339-361. doi:10.2307/146123

Rhodes RE, Janssen I, Bredin SSD, Warburton DER, Bauman A (2017) Physical activity: Health impact, prevalence, correlates and interventions. Psychol Health 32(8):942-975. doi:10.1080/08870446.2017.1325486

Sellers KF, Premeuax B (2021) Conway-Maxwell-Poisson regression models for dispersed count data. WIREs Comput Stat 13(6):e1533 doi: 10.1002/wics.1533

Slymen DJ, Ayala GX, Arredondo EM, Elder JP (2006) A demonstration of modelling count data with an application to physical activity. Epidemiol Perspect Innov 3:3. doi:10.1186/1742-5573-3-3

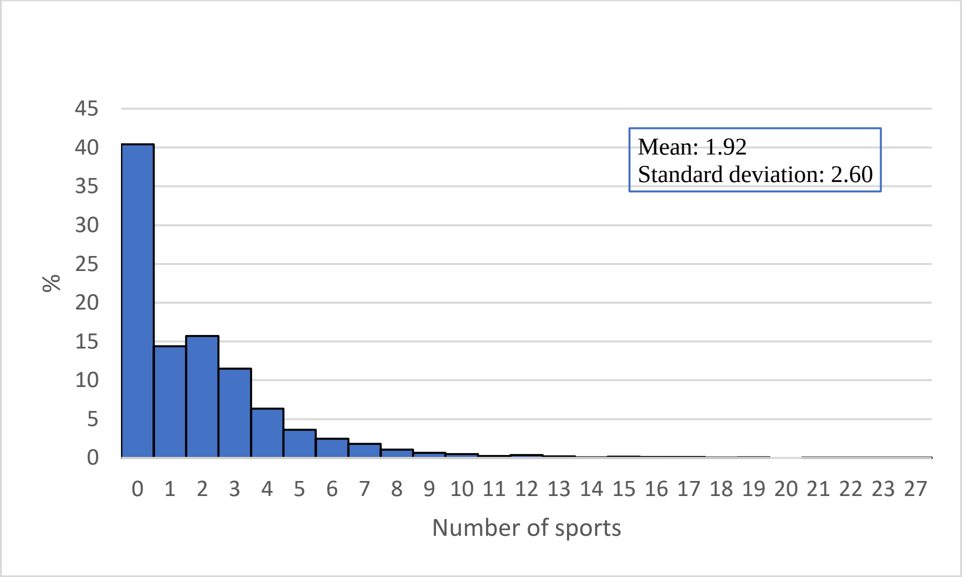
Sun J, Zhao X (2013) Analysis of Panel Count Data, Springer, New York.
doi: 10.1007/978-1-4614-8715-9

Tang W, He H, Tu XM (2023) Applied categorical and count data analysis, 2nd edn. CRC Press, Boca Raton. doi: 10.1201/9781003109815

WHO (2018) Global action plan on physical activity 2018-2030: More active people for a healthier world.
https://apps.who.int/iris/bitstream/handle/10665/272722/9789241514187-eng.pdf?ua=1.
Accessed 8 June 2023

Winkelmann R (2008) Econometric analysis of count data, 5th edn. Springer-Verlag, Berlin. doi:10.1007/978-3-540-78389-3

Figure 1: Histogram of the number of sports practised



Note: The values have been calculated using the corresponding weight factors.

Table 1: Descriptive statistics

Variables	**%**	**Sports participants (%)**	**Average number of sports by participant**
Gender			
Male	48.49	65.52	3.450
Female	51.51	54.01	2.969
Age			
<25	11.96	79.78	4.392
25-34	13.18	75.44	3.678
35-44	17.77	70.71	3.324
45-54	18.88	62.62	3.065
>54	38.22	41.13	2.273
Education			
Less than primary school	2.35	19.15	2.019
Primary school	12.63	31.84	2.233
Secondary school level 1	23.76	52.55	3.064
Secondary school level 2	28.02	64.77	3.433
University degree	33.25	73.65	3.339
Employment status			
Worker	48.84	68.42	3.356
Unemployed	11.73	57.77	2.883
Retired	23.56	36.95	2.062
Student	9.55	81.01	4.454
Housework	4.58	39.16	2.224
Other employment situation	1.75	66.48	3.547
Marital status			
Single	36.81	64.51	3.518
Married	58.69	56.98	2.995
Other	4.50	53.33	3.537
Health status			
Very good	24.86	75.51	3.563
Good	51.59	62.04	3.125
Fair	18.12	39.39	2.965
Bad	4.16	31.93	2.667
Very bad	1.28	27.43	1.757
Parents practising sport			
No	70.65	54.29	2.808
Yes	29.35	72.34	3.980
Nationality			
Spanish	91.00	59.63	3.203
Double nationality	3.27	59.90	3.707
Foreigner	5.72	58.82	3.309
Municipality			
Provincial capital	32.83	58.52	3.323
> 50,000 inhabitants	20.92	60.08	3.057
< 50.000 inhabitants	46.25	60.12	3.234

Note: The values have been calculated using the corresponding weights.

Table 2: *Stata* commands

Poisson:
poisson ndep i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.estfis i.pract_padres_new i.nac /// i.tram_munic
Negative Binomial:
nbreg ndep i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.estfis i.pract_padres_new i.nac /// i.tram_munic
Zero-Inflated Poisson:
zip ndep i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.pract_padres_new i.nac /// i.tram_munic, inflate(i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.estfis /// i.pract_padres_new i.nac i.tram_munic
Zero-Inflated Negative Binomial:
zinb ndep i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.pract_padres_new i.nac /// i.tram_munic, inflate(i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.estfis /// i.pract_padres_new i.nac i.tram_munic
Poisson hurdle model:
logit practica i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.estfis i.pract_padres_new i.nac /// i.tram_munic tpoisson ndep i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.pract_padres_new i.nac /// i.tram_munic tnbreg ndep i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.pract_padres_new i.nac /// i.tram_munic
Negative-Binomial hurdle model:
logit practica i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.estfis i.pract_padres_new i.nac /// i.tram_munic tnbreg ndep i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.pract_padres_new i.nac /// i.tram_munic
Ordered logit:
ologit ndep_o i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.estfis i.pract_padres_new i.nac /// i.tram_munic
Hurdle version of the ordered logit:
logit practica i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.estfis i.pract_padres_new i.nac /// i.tram_munic ologit ndep_o i.sexo i.g_edad i.estud_new i.sitlab i.sitpers_new i.pract_padres_new i.nac /// i.tram_munic if ndep>0

Note: The variable *practica* in the logit and in the Zero-Inflated models is a dummy variable equal to one when *ndep*>0 and equal to zero when *ndep*=0.

Table 3: Standard models

	Poisson	**NegBin**	**Ordered Logit**[1]
Gender (*Ref.: Male*)			
Female	-0.277***	-0.294***	-0.414***
Age (*Ref.: < 25*)			
25-34	-0.088**	-0.116	-0.124
35-44	-0.219***	-0.249***	-0.269*
45-54	-0.343***	-0.381***	-0.545***
>54	-0.631***	-0.673***	-0.915***
Education (*Ref.: Less than primary*)			
Primary school	0.355**	0.328*	0.311
Secondary school level 1	0.660***	0.642***	0.609**
Secondary school level 2	0.875***	0.892***	0.989***
University degree	1.046***	1.083***	1.395***
Employment status (*Ref.: Worker*)			
Unemployed	-0.162***	-0.147***	-0.239***
Retired	-0.342***	-0.315***	-0.405***
Student	0.172***	0.188**	0.443***
Housework	-0.322***	-0.290***	-0.336**
Other employment situation	0.162**	0.200*	0.201
Marital status (*Ref.: Single*)			
Married	-0.008	0.006	-0.000
Other	-0.103**	-0.136	-0.217
Health status (*Ref.: Very good*)			
Good	-0.074***	-0.091**	-0.217***
Fair	-0.248***	-0.314***	-0.642***
Bad	-0.528***	-0.601***	-0.927***
Very bad	-0.834***	-0.774***	-0.995***
Parents practising sport (*Ref.: No*)			
Yes	0.343***	0.344***	0.597***
Nationality (*Ref.: Spanish*)			
Double nationality	-0.020	0.036	-0.180
Foreigner	-0.067	-0.084	-0.237**
Municipality (*Ref.: Provincial capital*)			
> 50,000 inhabitants	-0.043	-0.020	-0.021
< 50.000 inhabitants	0.046**	0.072*	0.152***
Constant	0.299*	0.312	
α		0.766***	
Observations	5,151	5,151	5,151
log L	-10449	-9152	-8981

Notes: [1] The cut point estimates of the ordered Logit model are not reported.
*** p<0.01, ** p<0.05, * p<0.10

Tabla 4: Zero-inflated and hurdle (two-part) models

	Zero-inflated				**Hurdle (two-part)**			
	Poisson		**NegBin**		**Tr. Poisson**	**Tr. NegBin**	**Ordered**[1]	**Logit**[2]
	Number	**Inflate**	**Number**	**Inflate**	**Number**	**Number**	**Number**	**1st hurdle**
Gender (*Ref.: Male*)								
Female	-0.160***	0.435***	-0.219***	0.378***	-0.158***	-0.187***	-0.207***	-0.472***
Age (*Ref.: < 25*)								
25-34	-0.048	0.196	-0.055	0.330	-0.047	-0.059	-0.062	-0.199
35-44	-0.161***	0.295	-0.188**	0.336	-0.159***	-0.187**	-0.197	-0.347*
45-54	-0.233***	0.513**	-0.257***	0.748*	-0.232***	-0.273***	-0.434**	-0.569***
>54	-0.442***	0.738***	-0.499***	0.970**	-0.441***	-0.516***	-0.797***	-0.886***
Education (*Ref.: Less than primary*)								
Primary school	0.128	-0.288	0.180	-0.204	0.128	0.119	-0.076	0.340
Secondary school level 1	0.293	-0.535	0.356	-0.453	0.289	0.292	0.110	0.653**
Secondary school level 2	0.389*	-0.915***	0.485*	-0.934*	0.385*	0.423	0.378	1.033***
University degree	0.448**	-1.448***	0.571**	-1.711***	0.442**	0.491*	0.561	1.530***
Employment status (*Ref.: Worker*)								
Unemployed	-0.128***	0.139	-0.127**	0.200	-0.128***	-0.133**	-0.224**	-0.188*
Retired	-0.242***	0.189	-0.227***	0.265	-0.257***	-0.272***	-0.468***	-0.291***
Student	0.153***	-0.144	0.164**	-0.185	0.153***	0.175**	0.478***	0.192
Housework	-0.211**	0.118	-0.230*	0.110	-0.208**	-0.225*	-0.394*	-0.227
Other employment situation	0.141**	-0.028	0.137	-0.117	0.142**	0.183	0.237	0.089
Marital status (*Ref.: Single*)								
Married	-0.007	0.001	-0.020	-0.073	-0.008	-0.006	0.017	-0.004
Other	-0.073	0.249	-0.070	0.404	-0.068	-0.078	-0.081	-0.269*
Health status (*Ref.: Very good*)								
Good		0.411***		0.779***				-0.346***
Fair		1.037***		1.634***				-0.908***
Bad		1.296***		1.964***				-1.141***
Very bad		1.208***		1.671***				-1.055***

	Zero-inflated				Hurdle (two-part)			
	Poisson		**NegBin**		**Tr. Poisson**	**Tr. NegBin**	**Ordered**[1]	**Logit**[2]
	Number	**Inflate**	**Number**	**Inflate**	**Number**	**Number**	**Number**	**1st hurdle**
Parents practicing sport (*Ref.: No*)								
Yes	0.237***	-0.446***	0.282***	-0.426***	0.236***	0.263***	0.491***	0.521***
Nationality (*Ref.: Spanish*)								
Double nationality	0.093	0.348*	0.157	0.675**	0.087	0.141	0.153	-0.253
Foreigner	0.040	0.376**	0.029	0.601**	0.035	0.025	-0.029	-0.305**
Municipality (*Ref.: Provincial capital*)								
> 50,000 inhabitants	-0.088***	-0.189*	-0.105**	-0.428**	-0.085***	-0.090*	-0.196**	0.105
< 50.000 inhabitants	-0.011	-0.252***	-0.003	-0.383***	-0.011	-0.009	0.053	0.213***
Constant	0.946***	-0.857**	0.751***	-1.901***	0.950***	0.814***		0.571*
α			0.416***			0,422***		
Observations	5,151	5,151	5,151	5,151	3,150	3,150	3,150	5,151
log L[3]	-9494		-9020		-9493	-9019	-8947	

Notes: [1] The cut point estimates of the ordered Logit model are not reported.

[2] The Logit model of the first hurdle is the same for all the different versions of the hurdle model.

[3] The log likelihood of the Logit model in the two-part hurdle models is included in the value which appears in the second equation

*** p<0.01, ** p<0.05, * p<0.10

Table 5: Average of the adjusted probabilities for different values of *y* and the adjusted expected value of *y*

Prob.	Sample	Poisson	NegBin	Ordered Logit	Zero-infl. Poisson	Zero-infl. NegBin	Hurdle Poisson	Hurdle NegBin	Hurdle Ord. Logit
0	0.3885	0.2258	0.3593	0.3859	0.3886	0.3890	0.3885	0.3885	0.3885
1	0.1452	0.2562	0.2258	0.1437	0.1129	0.1686	0.1136	0.1696	0.1436
2	0.1633	0.2007	0.1388	0.1628	0.1388	0.1363	0.1388	0.1363	0.1611
3	0.1167	0.1356	0.0879	0.1179	0.1266	0.0989	0.1263	0.0986	0.1172
4	0.0660	0.0835	0.0573	0.0674	0.0954	0.0684	0.0952	0.0682	0.0673
5	0.0379	0.0479	0.0384	0.0387	0.0626	0.0462	0.0626	0.0461	0.0388
6	0.0239	0.0258	0.0262	0.0243	0.0370	0.0309	0.0370	0.0309	0.0244
7	0.0184	0.0131	0.0182	0.0187	0.0200	0.0205	0.0200	0.0206	0.0188
8	0.0109	0.0064	0.0129	0.0110	0.0100	0.0137	0.0100	0.0137	0.0110
9	0.0080	0.0029	0.0092	0.0080	0.0047	0.0091	0.0047	0.0091	0.0080
10	0.0052	0.0013	0.0066	0.0053	0.0020	0.0061	0.0021	0.0061	0.0053
11	0.0037	0.0005	0.0048	0.0037	0.0008	0.0040	0.0009	0.0041	0.0037
12	0.0035	0.0002	0.0036	0.0035	0.0003	0.0027	0.0003	0.0027	0.0035
13	0.0021	0.0001	0.0026	0.0021	0.0001	0.0018	0.0001	0.0018	0.0021
14	0.0012	0.0000	0.0020	0.0012	0.0000	0.0012	0.0000	0.0012	0.0012
15	0.0014	0.0000	0.0015	0.0014	0.0000	0.0008	0.0000	0.0008	0.0014
16	0.0016	0.0000	0.0011	0.0016	0.0000	0.0006	0.0000	0.0006	0.0016
17	0.0008	0.0000	0.0008	0.0008	0.0000	0.0004	0.0000	0.0004	0.0008
18	0.0002	0.0000	0.0006	0.0002	0.0000	0.0003	0.0000	0.0003	0.0002
19	0.0004	0.0000	0.0005	0.0004	0.0000	0.0002	0.0000	0.0002	0.0004
20	0.0002	0.0000	0.0004	0.0014	0.0000	0.0001	0.0000	0.0001	0.0014
E(*y*)	1.9851	1.9851	2.0003	2.0010	1.9847	1.9860	1.9842	1.9855	2.0010

Table 6: Marginal effects[1] of gender and health status (fair vs very good) on the probability of *n* and the expected value

	Poisson	NegBin	Ordered Logit	Zero infl. Poisson	Zero infl. NegBin	Hurdle Poisson	Hurdle NegBin	Truncated Ord. Logit
Gender: female vs male								
Pr(0)	0.0715	0.0666	0.0804	0.0916	0.0768	0.0925	0.0925	0.0925
Pr(1)	0.0406	0.0153	0.0041	0.0089	0.0092	0.0084	-0.0010	-0.0063
Pr(2)	0.0010	-0.0043	-0.0102	-0.0028	-0.0046	-0.0032	-0.0104	-0.0184
Pr(3)	-0.0209	-0.0103	-0.0185	-0.0152	-0.0115	-0.0154	-0.0141	-0.0202
Pr(4)	-0.0268	-0.0112	-0.0156	-0.0212	-0.0134	-0.0212	-0.0141	-0.0145
E(*y*)	-0.5462	-0.5829	-0.4472	-0.5259	-0.5408	-0.5258	-0.5386	-0.4149
Health status: fair vs very good								
Pr(0)	0.0658	0.0725	0.1289	0.1836	0.1792	0.1857	0.1857	0.1857
Pr(1)	0.0366	0.0161	0.0049	-0.0433	-0.0604	-0.0447	-0.0607	-0.0540
Pr(2)	-0.0017	-0.0058	-0.0193	-0.0468	-0.0435	-0.0478	-0.0445	-0.0517
Pr(3)	-0.0213	-0.0121	-0.0309	-0.0377	-0.0283	-0.0381	-0.0295	-0.0330
Pr(4)	-0.0251	-0.0127	-0.0246	-0.0253	-0.0177	-0.0253	-0.0188	-0.0175
E(*y*)	-0.4735	-0.5950	-0.6795	-0.5321	-0.4996	-0.5319	-0.5312	-0.5408

Note: [1] Average marginal effects for all the individuals in the sample