Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1804

The relevance of the specification assumptions when modelling the correlates of physical activity: an analysis across dimensions

Jaume Garcia-Villar and María José Suárez

November 2021

The relevance of the specification assumptions when modelling the correlates of physical activity: an analysis across dimensions

Jaume García
Departament d'Economia i Empresa
Universitat Pompeu Fabra
Barcelona School of Economics
ORCID ID: 0000-0002-8427-9541

María José Suárez
Departamento de Economía
Universidad de Oviedo
ORCID ID: 0000-0002-8152-1435

Abstract

There is a widespread economics literature on the determinants of sports participation and frequency, but the empirical evidence on the robustness of results to changes in the specification assumptions is still scarce. The goal of this paper is to contribute to fill this gap. To this end, we discuss and estimate different models – most of them previously applied in the literature – to check whether the econometric model, the functional form and the definition of physical activity (participation, time, frequency or intensity) condition the results. In particular, we study the probability of exercising, the frequency of participation in the previous week (days of practice), the number of minutes allocated and the intensity of exercise, with different econometric models and functional forms, using a data set from Mexico, where information about these four dimensions of practice of physical activity is available.

Keywords: sports participation, frequency, time, intensity, econometric modelling

JEL codes: Z29, C25, C52

Acknowledgements

We wish to thank the participants in the XI Conference of the European Sport Economics Association (ESEA), held in Gijón in May 2019, for their comments on an earlier version of this paper. Jaume García and María José Suárez wish to acknowledge financial support from projects ECO2017-83668-R and ECO2017-86402-C2-1-R, respectively. Jaume Garcia acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D ((CEX2019-000915-S).

The relevance of the specification assumptions when modelling the correlates of physical activity: an analysis across dimensions

Abstract

There is a widespread economics literature on the determinants of sports participation and frequency, but the empirical evidence on the robustness of results to changes in the specification assumptions is still scarce. The goal of this paper is to contribute to fill this gap. To this end, we discuss and estimate different models – most of them previously applied in the literature – to check whether the econometric model, the functional form and the definition of physical activity (participation, time, frequency or intensity) condition the results. In particular, we study the probability of exercising, the frequency of participation in the previous week (days of practice), the number of minutes allocated and the intensity of exercise, with different econometric models and functional forms, using a data set from Mexico, where information about these four dimensions of practice of physical activity is available.

Keywords: sports participation, frequency, time, intensity, econometric modelling

JEL codes: **Z29, C25, C52**

The relevance of the specification assumptions when modelling the correlates of physical activity: an analysis across dimensions

1. Introduction

The World Health Organization (WHO) Global Action Plan on Physical Activity 2018-2030 sets as its main target a 15% reduction in prevalence of insufficient physical activity in adults and adolescents by 2030. The policy actions to be taken by countries are explained not only by the health benefits but also because the objectives of this plan are aligned with the goals adopted by United Nations Member States in 2015 in connection with the 2030 Agenda for Sustainable Development. In this context, the study of the correlates of sports practice and exercise has received close attention to help to design and inform these actions, and nowadays the research on this topic in the economic literature is quite extensive in developed countries. Cabane and Lechner (2015) and Muñiz and Downward (2019) review the literature, from both a theoretical and an empirical point of view, as well as studies on the effects of physical activity on dimensions such as health, employment outcomes, education, well-being and social capital. Moreover, Rhodes et al. (2017) provide an overview of the scope of physical activity research through five topics: definition, health impact, physical activity levels, correlates and interventions to promote physical activity.

Although most of the sports economics literature is mainly empirical, those studies that include a theoretical framework generally follow the neoclassical approach. In particular, Becker's allocation-of-time model (Becker, 1965) has been the main starting point and the basis of the SLOTH framework (Cawley, 2004), which is the economic approach of reference in the majority of the literature. Becker focuses on the uses of time not allocated to work and emphasises the relevance of leisure time as an input in the production of household commodities. From this perspective, sports participation may be considered as a consumption good or as a leisure activity that has a direct effect on individual wellbeing, or it may be an

input in the production of health, health being an argument of the individual utility function. In this light, Downward and Rasciute (2010) point out that sports participation decisions should be connected with other leisure choices, and Humphreys and Ruseski (2011) present a model where the sports participation decision and the time spent on this activity are considered separate decisions and arguments of the utility function. Other theoretical approaches, sometimes called heterodox theories, underline the relevance of previous experience and social interactions in determining preferences and sports participation. [1]

Regarding empirical analyses, the economics literature on sports participation is quite heterogeneous in terms of the definition of the dependent variable, the type of survey used in the analysis and the econometric methodology, making it difficult to compare results. Sports practice has been defined as a dummy variable, an ordered variable, a continuous variable or a count variable, depending on the information. Moreover, the activities considered as sports or exercise vary among studies. Furthermore, the survey questions about physical activity refer to different time periods, with the corresponding implications for the characteristics of the participation variable to be explained. In addition, the econometric methods are quite diverse, largely conditioned by the nature of the dependent variable and the specific features of the available information.

Although some studies compare the estimation results for different dimensions of the dependent variable and others examine different econometric specifications for a given dimension (e.g. Meltzer and Jena, 2010; Dawson and Downward, 2011; Borgers et al., 2016), there is no overall discussion of the modelling of the different dimensions of physical activity. The aim of this research is to make a more comprehensive comparison of results from different econometric models and functional forms on the correlates of participation, time, frequency and intensity of physical activity. Additionally, we also contribute to the literature with a discussion about the functional form of the dependent variable, when continuous, by

applying the Box-Cox transformation. In our analysis, we use the 2015, 2016 and 2017 waves of the Mexican National Consumer Confidence Survey (*Encuesta Nacional sobre Confianza del Consumidor* or ENCO), which offers information about four dimensions of sports practice in the previous week (participation, time, frequency and intensity).

The remainder of the paper is structured as follows. The second section provides a summary of the economic literature on mass sports participation, highlighting the diversity of both the dependent variable definitions and the econometric models. The third section describes the data source, the variables and the empirical models applied. The fourth section discusses the results and the last section concludes.

2. Literature overview

Physical activity can be measured in different ways and this conditions the empirical specification. In this section, we briefly review the different dimensions of physical activity and discuss the econometric models applied in the literature. To some extent, this review complements and updates that of Downward and Rasciute (2010).

2.1. Dimensions of the dependent variable

When analysing sports practice or physical activity, several issues must be considered: Which activities should be included? How regularly is it done? How much time is allocated? How vigorously is it performed? What are the reasons for lack of participation? Some of these questions are related to the FITT principles that characterise physical activity, as mentioned by Rhodes et al. (2017): frequency (F), intensity (I), time (T) and type (T). The answers partly depend on the researcher's objective and the information available in the data source. Consequently, there is a wide variety of empirical approaches in the economics literature on sports participation.

The type of physical activity is related to the first question posed above (which activities should be included?). Economic studies usually focus on recreational sports practice

or exercise. However, physical activity done during daily tasks or transportation is sometimes considered. For instance, Buraimo et al. (2010) and Humphreys and Ruseski (2011) consider gardening, walking and/or cycling from one place to another along with sports.

Frequency (how regularly is it done?) and time (how much time is allocated?) are largely determined by the survey information. Frequency is measured as the number of days (times) in a certain period. Downward and Riordan (2007) and Muñiz et al. (2014) analyse the number of times that individuals played sports in four weeks; Borgers et al. (2016) define frequency as the number of times per week, whereas more qualitative information has been used by other researchers (Lera-López & Rapún-Gárate, 2007; Downward et al., 2014; Deelen et al., 2018, and García & Suárez, 2020).

Time (how much time is allocated?) is defined as the number of minutes or hours of sports or physical activity practice. García et al. (2011) consider hours in a day; Humphreys and Ruseski (2011) and Ruseski et al. (2011) study the hours or minutes allocated per week, whereas Eberth and Smith (2010) and Dawson and Downward (2011) analyse time in a four-week period, and Thibaut et al. (2017) study hours of practice over a year.

Intensity (how vigorously is it performed?) is the least analysed dimension of physical activity practice in the economics field. It may be measured either in a discrete way (moderate versus vigorous) or via metabolic equivalent tasks (METs), as in Meltzer and Jena (2010) and Garrues et al. (2017).

The last question posed at the beginning of this section (what are the reasons for lack of participation?) is an important issue in the analysis of sports/ physical activity practice because of the high proportion of people who do not exercise – so that the dependent variable takes the value 0 for them. In fact, many authors focus on the factors associated with the probability of exercising, therefore the variable is defined as binary (yes/no), and a minimum level of participation is established in some cases.[2]

2.2. Empirical methods

Logit and probit are the most common econometric models applied to the study of the probability of participation (Hovemann & Wicker, 2009; Breuer et al., 2011; Kokolakakis et al., 2012, and Dallmeyer et al., 2017). But the participation decision is often studied together with the frequency or the time of practice, to the extent that the observed values for these variables are zeros when the individuals have not exercised in the reference period. The main issue of discussion in the literature is how to model non-participation (e.g. Buraimo et al., 2010; Dawson and Downward, 2011; Humphreys & Ruseski, 2011, 2015; Thibout et al., 2017). There may different reasons for observing a zero, and each one is associated with a particular model. Humphreys and Ruseski (2011) distinguish between genuine and non-genuine zeros following the definition of Jones (2000), who states that a genuine zero corresponds to an actual choice of non-participation. According to the model specified by Humphreys and Ruseski (2011), where both the decision to practise sports and the time devoted to this activity are based on the assumption of utility maximisation, we could make an additional distinction between zeros associated with the participation decision (non-potential participants) and zeros linked to the choice of time, which are also the result of optimal decisions and can be defined as corner solutions.

Given the specific characteristics of sports participation, most zeros could be associated with non-potential participants, especially when the survey questions refer to regular sports practice. However, other surveys collect information about a particular period (e.g. last week or last month). Therefore, some zeros could come from potential participants who did not practise in the period due to their health conditions or professional commitments at the time of the interview, for instance.

Two-part models and sample selection models can be associated with zeros corresponding to non-potential participants, tobit models with zeros coming from corner solutions, and double-hurdle models with both types of zeros, nesting both the tobit and the sample selection model in its general formulation (Jones, 2000).

Moreover, it is worth mentioning the debate about the two-part versus the sample selection models, which has been very intense in the health economics literature (Duan et al., 1984; Madden, 2008) and which applies here. The sample selection model has been used in the limited dependent variable literature to deal with the estimation of an equation defined for positive observations only, but not as it was originally introduced, i.e. to solve a missing data (observability) problem. In the case of sports participation, there does not seem to be a missing data problem and researchers are not generally interested in the expected value of the dependent variable for those who are not potential participants. By contrast, the attractiveness of the two-part model is double. It is simple to estimate and it can be understood as a Taylor series approximation to a more general expression of the conditional expected value of time. However, some distributional assumptions (such as log-normality) or some orthogonality conditions (Duan, 1983) are required to work out the conditional and unconditional expected values of the dependent variable. But these distributional requirements are also present in the sample selection model and, in this case, exclusion restrictions are sometimes necessary to get precise estimates, given that the correction term is almost linear for most observations. In fact, the two-part model can be understood as if the reference population are potential participants.

All the models discussed above have been applied in the sports economic literature. Examples of tobit estimates can be found in Ruseski et al. (2011) and Thibaut et al. (2017). In other cases, researchers use the Heckman approach or sample selection model (e.g. Downward & Riordan, 2007; Buraimo et al., 2010; García et al., 2011)[3] or the two-part model

(Humphreys & Ruseski, 2011; Borgers et al., 2016; García & Suárez, 2020). Double-hurdle models are estimated in Buraimo et al. (2010) and Humphreys and Ruseski (2015).

In addition, count data models have been generally applied when studying the number of times/days of sports practice, and zero-inflated count data specifications are sometimes estimated to take into account the two possible sources of optimal zeros (e.g. Dawson & Downward, 2011; Muñiz et al., 2014). In fact, standard count data models are the equivalent to the tobit model when the dependent variable is a count, whereas the zero-inflated models are the counterpart of the double-hurdle models in their independent version.[4] When sports frequency is an ordinal (qualitative) variable, the standard ordered models or the zero-inflated versions, which are a combination of a discrete choice model and an ordered model, are often used (Lera-López & Rapún-Gárate, 2007; Downward et al., 2014; Downward & Rasciute, 2015).

3. Data and methods

In this section we first describe the data set used in the empirical analysis as well as the dependent and independent variables. Moreover, we provide some descriptive statistics of all the variables. Secondly, we briefly comment on the main characteristics of the econometric models applied in the empirical analysis.

3.1. Data and variables

The ENCO is a monthly database that is the result of a joint project between Mexico's Central Bank and the National Institute of Statistics and Geography. In our empirical analysis, we have pooled the data from November questionnaires conducted from 2015 to 2017, which offer information about sports participation. Given that the survey is a rotating panel, we have selected a random subsample of all households, so that no one appears twice in the sample. The selected sample consists of adults (people of 18 years of age or older) who answer the

survey module on sports practice and physical exercise, and provide information on all variables included.

The main advantage of this data set is that it includes questions about several dimensions of physical activity, which allows the use of different econometric models and different measures of sports practice to check the importance of the specification assumptions for the results obtained. In particular, the questions are: do you do sports (soccer, basketball, karate, …), exercise (walking, cycling, aerobics, …) or both? How many days last week? About how many minutes did you do per day last week? Was the intensity of this practice mainly moderate or vigorous?

Consequently, four alternative dependent variables have been used, all of them referring to the week prior to the survey: the participation decision (dummy variable equal to one if the person did physical activity); the minutes allocated to physical activity; the number of days per week allocated to this activity, and the intensity of participation (dummy variable equal to one if the individual did vigorous physical activity in the previous week and zero in the case of moderate physical activity). The descriptive statistics of these variables are included in the upper part of Table 1 for both the total sample and the subsample of participants. About 40% of individuals in the sample did some physical activity Among sports participants, they did physical activity on three days in the previous week and they allocated about three hours and a half to this activity on average. Around a fifth of participants did vigorous exercise.

Turning to the covariates, we have included personal and family characteristics, as well as income information. Specifically, gender is defined through a dummy variable equal to one for men (*male*), age is measured in years, marital status is included as a binary variable equal to one if married or in a cohabiting couple (*married*); family composition is measured as the number of children in the household under 12; education is measured through three

dummies for secondary, upper secondary and higher education respectively. Finally, employment status and two economic variables are also included: individual net weekly earnings[5] and net weekly earnings of other household members. Table 1 provides summary statistics of all variables.

Table 1. Descriptive statistics

	Total Sample		**Sports participants**	
	Mean	**St. Deviation**	**Mean**	**St. Deviation**
Participation	0.403	0.491	1.000	0.000
Time	85.447	150.814	211.843	172.098
Frequency	1.477	2.140	3.662	1.831
Intensity	-	-	0.213	0.409
Male	0.434	0.495	0.5083	0.500
Age	43.269	16.350	41.389	15.833
Married	0.574	0.495	0.544	0.498
#Children<12	0.630	0.917	0.563	0.882
Education				
Second. Ed.	0.247	0.431	0.206	0.405
Upper Sec. Ed.	0.182	0.386	0.192	0.395
Higher Ed.	0.348	0.476	0.444	0.497
Worker	0.637	0.481	0.633	0.482
Weekly Earnings (/1000)	1.065	1.398	1.217	1.720
Other earnings	1.349	1.829	1.477	2.163
# observations	4299		1734	

The proportion of men is higher among participants, unlike married status. In addition, people who did physical activity are somewhat younger, have a higher educational level, more earnings, and live in households with fewer children than those who did not engage in sports.

3.2.Methods

We have implemented different econometric specifications, taking into account the different dimensions of sports participation. First, a probit model is estimated to analyse the factors associated with the probability of doing physical activity, and it also constitutes the first stage of the two-part and sample selection models.[6]

Second, we estimate tobit, sample selection, two-part and double-hurdle models to simultaneously deal with physical activity participation and the time devoted to this activity. All these are two-equation models except the tobit model, for which the same equation explains both participation and the number of minutes devoted to physical activity.

The sample selection model (Heckman, 1979) can be understood as a simple way of overcoming the limitation of the tobit model of having just one equation to explain two dependent variables: participation and time. It assumes that the subsample of participants is not randomly selected from the population, so that there may be a sample selection problem, which is corrected by means of the specification of two equations (participation and time) that may be correlated. The Heckman approach is appropriate when the dependent variable is not always observed and the researcher is interested in the mean response of the population. But, in the case of estimating a sports participation model, the problem is not missing data, but zeros associated with non-participation. From an econometric point of view, Heckman's model is the same for dealing with a missing data problem or a problem of zeros, but the interpretation of the results is not exactly the same, because the unconditional expectation of the dependent variable varies depending on the problem. In the case of a problem of zeros, it is the product of the conditional (on being a participant) expectation of time, which includes a correction term, and the probability of being a participant. Thus, all the variables in both equations could potentially affect the unconditional expectation of time devoted to physical activity.[7]

The two-part models show some similarities with the sample selection model, since both assume a two-stage decision procedure. First, individuals choose whether to play sports, and those who opt to participate then decide how much time to allocate to that activity, and this amount is positive. In the case of two-part models, the two equations are independent, so that the likelihood function can be split into two components, each one associated with each dependent variable, allowing the separate estimation of each equation. The first stage is estimated via a probit (logit) model, whereas the second stage consists of the estimation of the time equation for the subsample of participants. In the time equation, the dependent variable is a positive random variable and a distribution accounting for this feature has to be specified (log-normal, Gompertz, truncated normal at zero, or gamma, among others).

The double-hurdle model is also a two-equation model in which two hurdles must be cleared in order to observe a positive value of the time variable. The first step consists of the potential participation decision and the second stage refers to the amount-of-time decision. The main difference with respect to the previous models is that here we may observe zeros among potential participants, because some of them may not have done exercise in the recorded period. Therefore, the double-hurdle model allows two types of zeros: those coming from people who would never participate (non-potential participants) and those who are potential participants but have not done sports in the period (corner solutions). In this case, the probability of non-participation is equal to the probability of either being a non-potential participant or being a potential participant but choosing zero minutes.

The maximum likelihood estimation of all these models requires some strong distributional assumptions. In particular, it is well known in the microeconometric literature that normality and homoscedasticity, two mostly imposed constraints, are necessary for the consistency of the maximum likelihood estimates (Amemiya, 1984). However, the normality assumption is violated whenever the dependent variable has a skewed distribution. Sometimes

this issue has been solved by using a logarithmic transformation of the dependent variable. But a more flexible approach is proposed by Yen and Jones (1996), when estimating a double-hurdle model, consisting of the following Box-Cox transformation (Box & Cox, 1964) for the dependent variable (*y*):

$$y_i^L = \frac{y_i^\theta - 1}{\theta} \quad for\ \theta > 0 \tag{1}$$

$$= \ln(y_i) \quad for\ \theta = 0$$

When applied to the context of the double-hurdle specification, this model nests both the standard version of the double-hurdle model (θ=1) and the two-part model with the dependent variable in logs (θ=0). This Box-Cox transformation of the dependent variable will also be used in the empirical exercise which follows.

When estimating the frequency of participation, we apply count data and ordered probit models, using the number of days that individuals have done physical activity as the dependent variable. In particular, we estimate Poisson, negative binomial (NB), ordered probit, two-part models (consisting of a probit in the first stage and a truncated Poisson or an ordered probit for the subsample of participants in the second stage) and zero-inflated Poisson (ZIP).

The Poisson, NB and ordered probit models are single-equation specifications that simultaneously explain participation and frequency, but they differ in the distributional assumptions. In fact, the Poisson model is very restrictive because of the equidispersion property (i.e. the expected value and the variance are equal) and the NB model allows for overdispersion, but both models impose a particular structure for the probabilities of the dependent variable taking the different values. In that sense the ordered model (McKelvey & Zavoina, 1975), which was not designed to deal with count data but with an ordered categorical variable, is more flexible in accommodating the structure of the probabilities because the cut-off points for the latent variable are estimated and not predetermined.

As mentioned above, the previous models are the "tobit" versions for count variables, since one equation explains both participation and frequency. It is possible to define two-equation models, similar to the two-part models and double-hurdle models considered for the time variable. In the case of the two-part models, there is an equation for the participation decision, estimated by applying a probit (logit) model, and the second stage consists of the estimation of the days of participation using a truncated (at zero) Poisson model. The models that are equivalent to the double-hurdle model are the so-called zero-inflated count data models (Poisson or NB), designed to deal with the empirical problem of "excess of zeros". The zero-inflated models assume independence between the two equations.

Regarding the intensity of sports practice, we apply a probit model with sample selection. In this model, there are two binary dependent variables, one for participation (practising physical activity) and another one for intensity, because of the limited information available in the survey about how vigorous the physical activity was. Intensity is only observed for participants. The random terms in both equations are allowed to be correlated.

4. Results

In this section we present the estimation results of several models for different dimensions of physical activity and compare the estimated coefficients. Given that the models are non-linear, we also compute and discuss the average marginal effects of the covariates on the probability of engaging in physical activity, on the expected value of time/frequency, and on the expected time/frequency conditioned to participation.

First, we will separately discuss the results for each dimension of physical activity (participation, time, frequency and intensity), emphasising the differences among specifications. Finally, we will comment on the main differences across dimensions.

4.1. Participation and time

Table 2 offers information about the estimated coefficients of the participation equation, which is also the first step in the two-part models, and the time equation. For the participation equation we report the probit estimates, whereas for the time dimension we provide the estimates of tobit, Heckman, two-part and double-hurdle models using the original time variable (minutes of exercise last week) and/or the logarithmic and Box-Cox transformations.

Table 2. Participation and time equations: coefficients

	Probit	Tobit	Heckman	Double-Hurdle		Two-part	Heckman	Tobit	Heckman	Two-part
	Partic.	**Time**	**Time**	**Pot. Partic.**	**Time**	**ln(Time)**	**ln(Time)**	**Box-Cox**	**Box-Cox**	**Box-Cox**
Male	0.334***	81.471***	1.851	0.185	64.215***	-0.022	-0.177***	95.220***	-0.120	-0.037
Age	-0.006***	-2.063***	-0.877***	-0.007	-1.301**	-0.004***	0.000	-3.399***	-0.006*	-0.009***
Married	-0.070*	-29.973***	-29.590***	0.327**	-60.575***	-0.090**	-0.057	-42.680***	-0.170**	-0.193**
#children<12	-0.081***	-26.554***	-14.538***	0.111	-34.036***	-0.053**	0.000	-41.979***	-0.081	-0.112**
Education										
Secondary ed.	0.102	18.075	-	0.081	-	-	-	-	-	-
Upper sec..ed.	0.264***	59.976***	-	0.565**	-	-	-	-	-	-
Higher ed.	0.469***	113.153***	-	1.594**	-	-	-	-	-	-
Worker	-0.324***	-73.651***	-9.202	-0.292	-76.685***	-0.111**	0.058	-101.661***	-0.118	-0.204**
Weekly earnings	0.067***	9.347**	-2.070	0.320**	7.543	-0.002	-0.045***	24.106***	-0.026	-0.005
Other earnings	0.024**	7.486***	4.059**	0.029	6.543**	0.017*	0.004	11.048***	-0.027	0.034*
Constant	-0.164	-15.874	278.799***	0.395	110.415***	5.390***	5.863	86.047***	8.815***	7.977
σ		288.580***	169.636***		270.265***	0.772***	0.968***	353.776***	1.523***	1.504***
ρ			-0.032	-0.115			-0.752***		-0.255	
θ								1.038***	0.132***	0.134***
Log L	-2759.56	-13735.78	-14121.18	-13723.43		-13558.85	-13549.74	-13678.42	-13544.67	-13545.20
AIC	5541.11	27495.56	28284.36	27488.87		27157.70	27141.48	27376.84	27133.34	27132.40

Notes: *** p<0.01, ** p<0.05, * p<0.1

σ is the standard deviation of the error term of the time equation.

ρ is the correlation coefficient between the error terms of the two-equations models.

θ is the parameter of the Box-Cox transformation.

Starting with participation, the probit and the first-stage Heckman estimates are practically the same, which is why the latter are not reported in the table.[8] All variables included have a significant effect in explaining participation. The likelihood of exercising is higher among males, young people and highly educated individuals. Being married and the number of children under 12 reduce the probability of participation, workers are less likely to engage in physical activity, and both own and others' household earnings increase participation. Notice that the results of the probit model are identical in terms of sign and significance to those of the tobit model, where one equation explains both participation and time. Nevertheless, the coefficients of the probit model do not seem to satisfy the proportionality between probit and tobit estimates when the tobit model is the correct model.[9]

By contrast, most of the variables lose their significance in the first stage of the double-hurdle model. It is worth recalling that, in this specification, zeros come from two sources: people who do not want to engage in physical activity (first stage of the double-hurdle model), and people who do but whose optimal participation is zero in the recorded period. Given that the participation equations are not equivalent in terms of the dependent variable, it may be that the variables that are not significant in the first hurdle (gender, age, children, employment status, and others' household earnings) would influence participation through their effect on the probability of being in a corner solution (time equation) rather than on the probability of being a potential participant. Only marital status, education and earnings remain significant and the sign of the marital status coefficient changes with respect to the probit model, which means that zeros for married people are mostly generated by the second hurdle.

Regarding the estimates of the time equations, there are considerable differences between the tobit model (second column of Table 2) and the sample selection and two-part models (third and sixth column of the table), and also between the last two models mentioned.

Neither gender nor weekly earnings is significant in the two-equation models, whereas employment status has a significant coefficient in the two-part model but not in the sample selection model. Additionally, on looking at the sample selection model with the time variable in logs (seventh column), results change substantially in terms of sign and significance when we compare the Heckman specification with the original time variable (third column). This could be explained by the negative and significant estimate of the covariance between the error terms of the participation and time equations when using the log-transformation versus the non-significant covariance when using the original time variable.

The estimates of the two-part model with the Box-Cox transformation, reported in the last column of Table 2, have the same sign and significance as those corresponding to the log-transformation. This is because the parameter θ, although significant, is very close to zero, which supports the log-transformation. This is not the case for the sample selection models, for which even with a θ coefficient close to zero, the results differ substantially from those obtained when the dependent variable is measured in logs (seventh and ninth columns of Table 2). This could be a consequence of the substantial change in the correlation coefficient - from -0.752, and significant, when using the log-transformation to -0.255, and not significant, when using the Box-Cox transformation, which affects the precision of the time estimates. It is also important to point out that the results from the tobit model with the Box-Cox transformation go in the same direction, in terms of sign and significance, as when using the original variable. In fact, the estimate of θ (1.038), although significantly different from 1, is very close to this value, so that the Box-Cox transformation does not imply a modification of the original variable, unlike the two-part and the sample selection models.

The double-hurdle model deserves special consideration. The coefficients of the second equation are highly significant, except that of the weekly earnings variable, and they have the same signs as in the tobit model. By contrast, the coefficients of the first hurdle

(potential participation equation) are quite imprecisely estimated, as we mentioned above. Thus, the second hurdle, which refers to the determinants of (potential) time, seems to dominate the generation of the zeros too. Although the Box-Cox transformation has also been applied to the double-hurdle model, no results are reported because it collapses into the sample selection model, i.e. all the zeros are generated in the first hurdle corresponding to non-potential participants.[10]

When looking at the goodness of fit of the different models by means of the Akaike information criterion (AIC), we can conclude that the Box-Cox transformation implies a substantial improvement with respect to the models in which the original variable is not transformed. The largest values of the AIC (the best fit) are associated with the two-part model, although they are very similar to the sample selection model, since the correlation between the error terms is not significantly different from zero. The two-equation models with the Box-Cox transformation clearly outperform the tobit version with substantially different implications about how the zeros are generated.

It is worth paying attention to the models in which the time variable is not transformed, because this is the most common specification in the literature (columns 1-5 of Table 2). The double-hurdle model outperforms the tobit and the sample selection models, which are nested in the former, and the tobit model has a better fit than the sample selection model. This evidence clearly illustrates how important the choice of the functional form for the dependent variable is, both in terms of the selection of the "best" model and in terms of the implications for the behaviour of the individuals, i.e. how the zeros are generated.

As mentioned above, all these models are highly non-linear, therefore the coefficients are not informative about the size of the effects of the explanatory variables or, in some cases, about the sign of the effects either. Thus, it is important to compute the marginal effects,

which are different for each individual since they depend on the values of the explanatory variables.

In Table 3 we report the mean values of the marginal effects calculated for all individuals. In particular, we compute the change in the probability of practising physical activity ($\partial Pr(y>0) / \partial x_j$), in the unconditional expected time ($\partial E(y) / \partial x_j$) and in the expected value of time conditional on being positive ($\partial E(y/y>0)/\partial x_j$). The marginal effects reported correspond to the tobit, sample selection and double-hurdle models included in the first five columns of Table 2. We have chosen these models because in all of them the dependent variable (time) is not transformed, as is usual in the empirical literature on the topic. We also report the marginal effects for the two-part models with both log and the Box-Cox transformations.[11] The sample selection model with the Box-Cox was discarded because it was statistically equivalent to the two-part model, and the tobit versions were also discarded because the evidence presented above does not support this particular specification.

Focusing on the two-part model with the Box-Cox transformation, which was the preferred specification (last column in Table 2), the interpretation of the marginal effects is as follows. Being a male, compared to a female, increases the probability of practising physical activity by 12.4 percentage points on average, increases by 24.1 the number of minutes per week devoted to sports practice, and decreases by 3.7 minutes the expected number of minutes conditional on practising physical activity. For the quantitative variables such as age or earnings, the marginal effects measure approximately changes in the corresponding covariate when the explanatory variable increases by one unit.

Table 3. Participation and time equations: marginal effects

		Tobit	Heckman	Double-hurdle	Two-part	Two-part
		Time	Time	Time	ln(Time)	Box-Cox
Male	$\partial Pr(y>0)/\partial x_j$	0.106	0.124	0.096	0.124	0.124
	$\partial E(y)/\partial x_j$	33.136	27.168	29.905	24.573	24.137
	$\partial E(y/y>0)/\partial x_j$	25.897	3.081	23.453	-4.621	-3.661
Age	$\partial Pr(y>0)/\partial x_j$	-0.003	-0.002	-0.002	-0.002	-0.002
	$\partial E(y)/\partial x_j$	-0.825	-0.858	-0.675	-0.916	-0.838
	$\partial E(y/y>0)/\partial x_j$	-0.650	-0.901	-0.486	-0.882	-0.852
Married	$\partial Pr(y>0)/\partial x_j$	-0.039	-0.026	-0.038	-0.026	-0.026
	$\partial E(y)/\partial x_j$	-12.073	-17.563	-16.518	-13.517	-13.172
	$\partial E(y/y>0)/\partial x_j$	-9.473	-29.851	-20.260	-19.342	-18.968
#children<12	$\partial Pr(y>0)/\partial x_j$	-0.034	-0.030	-0.029	-0.030	-0.030
	$\partial E(y)/\partial x_j$	-10.620	-12.140	-10.977	-10.968	-10.598
	$\partial E(y/y>0)/\partial x_j$	-8.361	-14.836	-11.654	-11.351	-10.966
Education						
Secondary ed.	$\partial Pr(y>0)/\partial x_j$	0.022	0.036	0.013	0.036	0.036
	$\partial E(y)/\partial x_j$	6.144	7.552	2.936	7.734	7.486
	$\partial E(y/y>0)/\partial x_j$	5.135	0.393	0.444	-	-
Upper sec.	$\partial Pr(y>0)/\partial x_j$	0.076	0.096	0.086	0.096	0.096
	$\partial E(y)/\partial x_j$	21.979	20.306	19.396	20.588	19.927
	$\partial E(y/y>0)/\partial x_j$	17.811	1.004	2.956	-	-
Higher ed.	$\partial Pr(y>0)/\partial x_j$	0.147	0.175	0.185	0.175	0.175
	$\partial E(y)/\partial x_j$	45.401	37.165	41.274	37.447	36.246
	$\partial E(y/y>0)/\partial x)$	35.582	1.742	6.497	-	-
Worker	$\partial Pr(y>0)/\partial x_j$	-0.095	-0.118	-0.122	-0.118	-0.118
	$\partial E(y)/\partial x_j$	-30.480	-29.177	-37.369	-35.940	-33.468
	$\partial E(y/y>0)/\partial x_j)$	-23.750	-10.381	-28.271	-23.938	-20.243
Weekly earni	$\partial Pr(y>0)/\partial x_j$	0.012	0.025	0.042	0.025	0.025
	$\partial E(y)/\partial x_j$	3.739	4.357	10.480	5.019	4.851
	$\partial E(y/y>0)/\partial x_j$	2.943	-1.823	3.865	-0.523	-0.523
Other earnin	$\partial Pr(y>0)/\partial x_j$	0.010	0.009	0.011	0.009	0.009
	$\partial E(y)/\partial x_j$	2.994	3.501	3.295	3.325	3.182
	$\partial E(y/y>0)/\partial x_j$	2.357	4.148	2.429	3.523	3.328

As reported in Table 3, there are no substantial differences in the effects of the independent variables on the probability of being a participant across specifications. Regarding time, we can observe differences in some marginal effects depending on whether the variable is transformed or not. This is the case for gender: the marginal effect on the conditional expectation is small and not significant when the Box-Cox transformation is used, but it is positive and highly significant when we consider the tobit model or the double-hurdle

model. This is a consequence of the significance of the gender variable in the second equation of the different models. The marginal effects of gender on the unconditional expectation are much more similar, but they are higher in the models without transformation of the dependent variable. In the case of weekly earnings, different signs of the marginal effects on the conditional expectation are found depending on the transformation of the dependent variable. For the remaining covariates, the main differences are in the size of the effects.

4.2. Frequency

Table 4 provides information about the coefficients of the models estimated for the number of days per week that people do physical activity. Given that this is a count variable, we have applied standard versions of count data models (Poisson and NB). Moreover, since the dependent variable can only take eight values (0-7), we have also estimated a standard ordered probit, treating the number of days as an ordered categorical variable. Finally, we estimate the two-part versions of these models, with a probit model for the participation equation, as well as the zero-inflated versions.

In the standard Poisson, NB and ordered probit, we find that the participation equation dominates the results again in terms of sign and significance of the coefficients. Moreover, the coefficients of these three models have the same sign as the probit (and tobit) model in Table 2, but the precision of the estimates of the NB model is much lower than that of the Poisson, as usual. Notice that there is a huge improvement in the log-likelihood when we allow for overdispersion and the coefficient (α) for the parameterisation of the variance in terms of the expected value is highly significant.[12] To some extent, this is reflected neither in the coefficients nor in the marginal effects on the expected number of days, as we will see later. Finally, the ordered probit model provides a much better fit because it imposes a much more flexible structure for the probabilities of the different values of the dependent variable.

Table 4. Frequency and intensity equations: coefficients

	FREQUENCY								INTENSITY	
	Poisson	Negative Binomial	Ordered Probit[1]	Two part (2nd step)			ZIP		Heckprobit	
	#days	#days	#days	Truncated Poisson	Trunc. NegBin (1-7)	Ordered Probit[1]	Pot. Partic.	#days	Partic.	Intensity (strong)
Male	0.204***	0.195***	0.229***	-0.103***	-0.202***	-0.228***	0.353***	-0.101***	0.334***	0.429***
Age	-0.002**	-0.003	-0.003**	0.004***	0.008***	0.008***	-0.007***	0.004***	-0.006***	-0.027***
Married	-0.073***	-0.063	-0.064*	-0.047*	-0.100*	-0.093*	-0.066	-0.050*	-0.070*	-0.040
#children<12	-0.095***	-0.096***	-0.073***	-0.024	-0.043	-0.053*	-0.079***	-0.024	-0.081***	-0.021
Education										
Secondary ed.	0.104**	0.086	0.088	-	-	-	0.101	-	0.102	-
Upper sec.ed.	0.261***	0.238**	0.220***	-	-	-	0.209***	-	0.264***	-
Higher ed.	0.495***	0.477***	0.415***	-	-	-	0.478***	-	0.469***	-
Worker	-0.341***	-0.348***	-0.283***	-0.111***	-0.223***	-0.250***	-0.325***	-0.113***	-0.324***	0.233**
Weekly earn.	0.023***	0.032	0.029*	-0.008	-0.012	-0.011	0.074***	-0.008	0.067***	-0.005
Other earn.	0.025***	0.023	0.023**	0.009	0.020	0.022*	0.024**	0.009	0.024**	0.015
Constant	0.372***	0.406***		1.263***	1.416***		-0.128	1.266***	-0.164	-0.0889
ρ										-0.0382
a		2.735***			-1.700***					
Log L	-8943.10	-6779.30	-6000.23	-6123.12	-6020.68	-5910.49	-6123.32		-3557.32	
AIC	17908.20	13582.60	12034.46	12284.24	12081.36	11868.98	12284.64		7154.64	

Notes: *** p<0.01, ** p<0.05, * p<0.1

[1] No constant term is reported for the ordered model since (6-7) cut-off points are estimated to define the intervals associated with each category.

ρ is the correlation coefficient between the error terms of the two equations.

α is the overdispersion parameter of the Negative Binomial model.

We also estimated the two-part versions of the standard models, where the number of days only takes positive values in the second equation (truncated model). Additionally, since the frequency variable refers to the week, we also take into account the upper truncation in 7. In Table 4 we report the estimates for the (left) truncated version of the Poisson model and the double-truncation version (values from 1 to 7) of the NB model.[13]

There is a substantial improvement in the value of the log-likelihood function when the two-part versions are estimated, in particular for the count data models. But the ordered model still has a better fit. As in the two-equation models for time, this huge improvement in the fit, compared to the standard versions, translates into different effects of the covariates in both equations.

In particular, the number of children younger than 12 and own earnings lose their significance in the second step of most two-part models. However, employment status is significant in all cases. According to the probit estimates of the first equation (first column of Table 2) workers are less likely to participate, and those who do, allocate fewer days to sports practice. Regarding the rest of the covariates, there are differences in the results depending on the specification. As in the analysis of time, when participation and frequency are modelled as separate decisions, some variables have a different effect on the probability of doing sports and on the number of days. Specifically, men and young people are more likely to do sports, but they practise it fewer days a week than women and older people respectively; and own and others' earnings do not significantly affect frequency, but they increase participation. Even when we focus on the second stage of the two-part models, some differences arise depending on the specification assumptions. For example, the number of children reduces the frequency of sports, according to the ordered probit model, but it is not significant in other specifications; and earnings of other household members increase the frequency of practice in some cases.

Turning now to the zero-inflated models, the zero-inflated NB model did not converge, for the same reasons pointed out above when discussing the two-part models. On the other hand, the zero-inflated version of the ordered probit is not reported because it converges to the two-part model, i.e. all the zeros are generated in the first equation.[14] The fit of the zero-inflated Poisson model, reported in Table 4, is much better than that of the standard model. This is because allowing the zeros to be generated by a different model matters. When we look at the coefficients, the estimates of the potential participation equation are very similar to those of the participation equation in Table 2.

According to the goodness of fit, we can conclude that a two-equation model is the most appropriate for the analysis of frequency, as in the case of time. Additionally, the two-part structure where the zeros are treated as corresponding to non-potential participants seems to be more appropriate, and the ordered models capture the empirical frequency patterns observed for this data set better than the standard count data specifications.

Given the non-linear nature of the models, it is advisable to compute the marginal effects. In Table 5 we report the same type of marginal effects that we calculated in the previous subsection for the time variable, and the interpretation is similar, except that now the variable under study is the number of days of sports practice per week.

Comparing one-equation versus two-equation models (i.e. two-part and zero-inflated specifications), the most notable difference is that those variables whose coefficient has different sign and/or significance in the two equations in the latter case – gender, age and weekly earnings – have a marginal effect on the conditional expectation which differs from that of the one-equation models.

Table 5. Frequency and intensity: marginal effects

		FREQUENCY							INTENSITY
		Poisson	Neg. Binomial	Ord. Probit	Two part models			ZIP	Heckprobit
					Trunc- Poisson	Trunc. Poisson (1-7)	Ord. Probit		Intensity (strong)
Male	$\partial Pr(y>0)/\partial x_j$	0.066	0.032	0.086	0.124	0.124	0.124	0.124	0.124
	$\partial E(y)/\partial x_j$	0.305	0.292	0.406	0.317	0.320	0.296	0.318	0.110
	$\partial E(y/y>0)/\partial x_j$	0.227	0.415	0.223	-0.339	-0.331	-0.388	-0.333	0.110
Age	$\partial Pr(y>0)/\partial x_j$	-0.001	-0.000	-0.001	-0.002	-0.002	-0.002	-0.002	-0.002
	$\partial E(y)/\partial x_j$	-0.003	-0.004	-0.005	-0.004	-0.003	-0.004	-0.004	-0.007
	$\partial E(y/y>0)/\partial x_j$	-0.003	-0.006	-0.003	0.012	0.013	0.013	0.012	-0.007
Married	$\partial Pr(y>0)/\partial x_j$	-0.024	-0.010	-0.024	-0.026	-0.026	-0.026	-0.026	-0.026
	$\partial E(y)/\partial x_j$	-0.109	-0.093	-0.113	-0.158	-0.163	-0.160	-0.161	-0.010
	$\partial E(y/y>0)/\partial x_j$	-0.081	-0.132	-0.062	-0.156	-0.165	-0.157	-0.165	-0.010
#children<12	$\partial Pr(y>0)/\partial x_j$	-0.031	-0.016	-0.027	-0.030	-0.030	-0.030	-0.030	-0.030
	$\partial E(y)/\partial x_j$	-0.140	-0.142	-0.128	-0.141	-0.139	-0.145	-0.141	-0.005
	$\partial E(y/y>0)/\partial x_j$	-0.104	-0.202	-0.071	-0.080	-0.074	-0.090	-0.081	-0.006
Education									
Secondary ed.	$\partial Pr(y>0)/\partial x_j$	0.037	0.014	0.032	0.036	0.036	0.036	0.036	0.036
	$\partial E(y)/\partial x_j$	0.124	0.103	0.137	0.133	0.133	0.133	0.134	-
	$\partial E(y/y>0)/\partial x_j$	0.086	0.149	0.079	-	-	-	0.000	0.001
Upper sec.	$\partial Pr(y>0)/\partial x_j$	0.091	0.039	0.081	0.096	0.096	0.096	0.097	0.097
	$\partial E(y)/\partial x_j$	0.337	0.308	0.360	0.356	0.356	0.355	0.358	-
	$\partial E(y/y>0)/\partial x_j$	0.239	0.444	0.204	-	-	-	0.000	0.002
Higher ed.	$\partial Pr(y>0)/\partial x_j$	0.164	0.077	0.157	0.175	0.175	0.175	0.174	0.175
	$\partial E(y)/\partial x_j$	0.722	0.700	0.731	0.648	0.648	0.648	0.645	-
	$\partial E(y/y>0)/\partial x)$	0.532	0.997	0.402	-	-	-	0.000	0.003
Worker	$\partial Pr(y>0)/\partial x_j$	-0.109	-0.056	-0.106	-0.118	-0.118	-0.118	-0.122	-0.118
	$\partial E(y)/\partial x_j$	-0.530	-0.542	-0.508	-0.596	-0.599	-0.624	-0.611	0.056
	$\partial E(y/y>0)/\partial x)$	-0.399	-0.766	-0.278	-0.375	-0.373	-0.430	-0.379	0.053
Weekly earn.	$\partial Pr(y>0)/\partial x_j$	0.008	0.005	0.011	0.025	0.025	0.025	0.027	0.025
	$\partial E(y)/\partial x_j$	0.035	0.047	0.050	0.080	0.080	0.083	0.088	-0.001
	$\partial E(y/y>0)/\partial x_j$	0.026	0.067	0.028	-0.026	-0.026	-0.018	-0.025	-0.001
Other earn	$\partial Pr(y>0)/\partial x_j$	0.008	0.004	0.009	0.009	0.009	0.009	0.009	0.009
	$\partial E(y)/\partial x_j$	0.036	0.033	0.040	0.044	0.047	0.047	0.044	0.004
	$\partial E(y/y>0)/\partial x_j$	0.027	0.047	0.022	0.029	0.036	0.037	0.029	0.004

Notes: *** p<0.01, ** p<0.05, * p<0.1. In the Heckprobit model $\partial E(y)/\partial x_j$ corresponds to $\partial Pr(str=1)/\partial x_j$ and $\partial E(y/y>0)/\partial x_j$ corresponds to $\partial Pr(str=1/y>0)/\partial x_j$.

Moreover, the marginal effect on the unconditional expectation is very much dominated by the effect on the probability of practising sports. But even in the cases where the sign of the marginal effect on the unconditional and conditional expectations are the same, there are some considerable differences in magnitude. This happens, for instance, with the marital status variable, whose negative effect is higher in absolute value in the two-equation models.

There are no substantial differences among the marginal effects of the two-equation models reported in Table 5. This could be surprising since the fit of the two-part ordered model seems to be much better than the rest, but we can better appreciate the different performance of these models when comparing the adjusted probabilities with the sample frequencies. In the case of the two-part ordered probit model, the average of the adjusted probabilities for each value of the dependent variable (from 0 to 7) is almost equal to the observed frequencies, unlike what happens with the count data models. In particular, the average adjusted probability for 1 day is 0.328 for the Poisson model and 0.161 for the NB model, when the sample frequency is 0.046. In the case of 7 days, these numbers are 0.002, 0.013 and 0.051 respectively. Given the specific features of these models, much more attention should be paid to the probabilities instead of focusing only on the expected values.

4.3.Intensity

In the analysis of intensity of practice, the only information available is whether it is moderate or vigorous, so that we estimated a probit model with sample selection (or Heckman probit), which allows for correlation between the error terms of the participation and intensity equations. The last two columns (Heckprobit) of Table 4 offer information about the estimated coefficients of the two equations: whether the individual practises sports and whether physical activity is vigorous or not among participants. Since the estimate of the correlation coefficient is not significantly different from zero, the first stage results are similar

to the ones included in the first column of Table 2. This means that participation and intensity could be modelled as independent decisions. The only relevant correlates of the intensity of practice are age, employment status and gender and they have the expected sign: males, workers, and young people are more likely to engage in vigorous physical activity. But the effect of these variables is just the opposite of what we found for frequency in the second step of two-equation models, and almost the same in comparison with two-equation models for time. This highlights the point that results vary depending on the dimension of sports practice considered.

In the last column of Table 5, we report the marginal effects for the intensity model. Males have an average probability of practising sports which is 12.4 percentage points (pp) higher than that for females, an unconditional probability of doing vigorous exercise which is 11 pp higher, and a conditional probability which is 10.9 pp higher.

4.4 Comparison across dimensions

When comparing different dimensions of physical activity, we find some interesting results. The effect of gender is striking. Males are more likely to participate regardless of the specification, and with more intensity. Its coefficient on time is either not significant or positive, but its coefficient on the number of days is negative in all the two-equation specifications. The research by Humphreys and Ruseski (2009) also reveals a different sign for the coefficient of gender in participation and frequency equations.

Another variable with a different effect depending on the dimension of physical activity is age. It decreases the probability of doing exercise, the intensity and the minutes allocated to that activity. However, its effect on the number of days conditional to participation is positive. Therefore, we may conclude that older people are less likely to exercise and those who participate allocate less time in total, but a greater number of days per

week. Meltzer and Jena (2010) obtain differences in the influence of age depending on the explained variable too.

Workers are less likely to participate and they allocate less time – when significant – and fewer days to physical activity, but their intensity of practice is greater than the rest. This result is in line with the hypothesis developed by Meltzer and Jena (2010) that a higher opportunity cost of time increases the intensity of physical activity.

All in all, two-equation models seem to perform better to explain participation and the different dimensions of physical activity. Although we have found some different effects of the covariates depending on the dimension considered (in particular between time and frequency), they are not as important as could be expected because of the weekly time interval considered. We would expect to find larger discrepancies if a longer reference period were taken. In fact, the correlation between the time and frequency variables in this survey is very high (0.59 when considering those practising physical activity at least once a week).

5. Conclusions

The goal of this paper is to contribute to the existing literature on the correlates of sports participation by making a broad comparative analysis of different econometric methodologies and dimensions of physical activity. Moreover, we also deal with the issue of functional form and apply the Box-Cox transformation, which allows for a flexible transformation of the dependent variable. For these purposes, we use an adult sample from the 2015-2017 waves of the Mexican National Consumer Confidence Survey (*Encuesta Nacional sobre Confianza del Consumidor* or ENCO).

We study four different dimensions of physical activity: the probability of participation; the number of minutes and the number of days a week allocated to this activity; and the degree of intensity (moderate or vigorous). Several specifications have been estimated depending on the nature of the dependent variable. A probit has been applied to estimate the

probability of participation. Regarding time, tobit, Heckman, double-hurdle and several two-part models have been estimated. In the case of frequency, Poisson, NB, ordered probit, ZIP and some two-part models have been tried. Finally, a probit model with selection has been applied to explain intensity of practice.

The results show that the effects of some covariates may differ depending on the specification within each dimension of physical activity – specifically time and frequency. In particular, the models that separate participation and time/frequency reveal the uneven effect of some covariates on each decision. There are also substantial disparities in the effect of some variables across dimensions. On this point, it is worth mentioning the opposite effect of gender and age on time and frequency. In general, the AIC benefits those specifications that separate participation and time/frequency.

Therefore, when studying the correlates of sports practice it is important to consider which is the variable of interest (time, frequency, intensity) and the reasons for not engaging in this activity (whether there may be corner solutions, infrequency of practice or lack of interest) because the appropriate model varies depending on these issues. It is also advisable to make robustness checks of the specification and the covariates included.

One of the limitations of our study is that the database does not offer information about some covariates often considered in the literature, such as sports facilities. Another future extension of this work would be to replicate the analysis with data from other countries, to check whether the conclusions are maintained, and with a more detailed measure of the intensity dimension.

Footnotes

[1] See Cabane and Lechner (2015) for a summary of the main theoretical explanations for involving in physical activity.

[2] In this respect, it is worth mentioning the paper by Wicker et al. (2017), who analyse different definitions of physical activity depending on the degree of compliance with the WHO guidelines (WHO, 2010).

[3] Eberth and Smith (2010) estimate the sample selection model using flexible parametric forms based on a copula approach, in which no normality assumptions are required in order to define the joint cumulative distribution function of both dependent variables.

[4] As mentioned in Jones (2000), sometimes hurdle and two-part models are used as synonymous terms in the count data literature, but they are not.

[5] We estimated an earnings equation with the subsample of individuals who offered information about this variable to impute earnings for those individuals who did not report it, taking into account potential sample selection problems.

[6] A logit model could also be estimated, but the probit estimates are reported since the normality assumption of the errors is assumed in most of the two-equation models, and the logit and probit models do not differ much unless there are many observations in the tails of the distributions.

[7] In fact, the presence of this correction term in the time equation may justify the empirical findings of some imprecise estimates of the coefficients of this equation and the consideration of exclusion restrictions, though they are not strictly necessary.

[8] The sample selection model has been estimated by maximum likelihood and this explains why the participation equation estimates are not numerically the same as in the probit model.

[9] An intuitive (visual) test to for the appropriateness of the tobit specification is to check whether the coefficients of the probit are those of the tobit model divided by the standard deviation of the error term.

[10] This is because, with the Box-Cox transformation, the condition associated with being a potential participant depends on $(1/\theta)$, where θ is the parameter of the Box-Cox transformation. Since the estimate of this parameter is very small (around 0.13), it has a huge influence for all the observations on the probability of being a potential (or a non-potential) participant.

[11] The calculation of the marginal effects on $\partial E(y/y>0)/\partial x_j$ with the Box-Cox transformation would require numerical integration. Instead, we use the proposal by Abrevaya (2002) based on a flexible estimator (smearing estimator) proposed by Duan (1983).

[12] The NB model estimated is the Type II version, where $var(y)=E(y)+\alpha[E(y)]^2$. On the other hand, the presence of overdispersion is evident when we look at the sample mean and the sample variance of the dependent variable (1.48 and 4.58 respectively).

[13] The NB version of the truncated model did not converge because there was no improvement in the log-likelihood compared to the Poisson model, and the sample mean and the sample variance showed underdispersion, which cannot be accounted for by an NB model. The Poisson version of the double-truncation model is not reported because, according to the estimates of the NB version, it is a clear case of overdispersion ($\alpha = 1.70$).

[14] In fact, when we look at the estimates of the zero-inflated ordered probit model the first cut-off point estimate is -5.847. This means that the probability of a zero being generated in the second equation is almost negligible for any individual in the sample.

Declarations of interest: None

References

Abrevaya, J. (2002). Computing marginal effects in the Box-Cox model. *Econometric Reviews, 21*(3), 383-393.

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics, 24*(1-2), 3-61.

Becker, G.S. (1965). A theory of the allocation of time. *The Economic Journal, 75*(299), 493-517.

Borgers, J., Breedveld, K. Tiessen-Raaphorst, A., Thibaut, E., Vandermeerschen, H., Vos, S., & Scheerder, J. (2016). A study on the frequency of participation and time spent on sport in different organisational settings. *European Sport Management Quarterly, 16*(5), 635-654.

Box, G.E.P., & Cox, D.R. (1964). An analysis of transformations (with Discussion). *Journal of the Royal. Statistical Society (Series B), 26*(2), 211-252.

Breuer, C., Hallmann, K., & Wicker, P. (2011). Determinants of sport participation in different sports. *Managing Leisure, 16*(4), 269-286.

Buraimo, B., Humphreys, B.R., & Simmons, R. (2010). Participation and engagement in sport: A double hurdle approach for the United Kingdom. *The Selected Works of Dr Babatunde Buraimo.* Retrieved from: http://works.bepress.com/babatunde_buraimo/3/.

Cabane, C., & Lechner, M. (2015). Physical activity of adults: A survey of correlates, determinants and effects. *Journal of Economics and Statistics, 235*(4-5), 367-402.

Cawley, J. (2004). An economic framework for understanding physical activity and eating behaviors. *American Journal of Preventive Medicine*, 27(3S), 117-125.

Dallmeyer, S., Wicker, P., & Breuer, C. (2017). Public expenditure and sport participation: An examination of direct, spillover, and substitution effects. *International Journal of Sport Finance, 12*(3), 244-264.

Dawson, P., & Downward, P. (2011). Participation, spectatorship and media coverage in sport: some initial insights. In W.Andreff (Ed.), *Contemporary issues in sports economics: participation and professional team sports* (pp. 15-42). Cheltenham: Edward Elgar.

Deelen, I., Ettema, D., & Kamphuis, C.B.M. (2018). Sports participation in sport clubs, gyms or public spaces: How users of different sports settings differ in their motivations, goals, and sports frequency. *PLoS ONE, 13*(10), e0205198.

Downward, P., Lera-López, F., & Rasciute, S. (2014). The correlates of sports participation in Europe. *European Journal of Sport Science, 14*(6), 592-602.

Downward, P., & Rasciute, S. (2010). The relative demand for sports and leisure in England. *European Sport Management Quarterly, 10*(2), 189-224.

Downward, P., & Rasciute, S. (2015). Exploring the covariates of sport participation for health: an analysis of males and females in England. *Journal of Sports Sciences*, *33*(1), 67-76.Downward, P., & Riordan, J. (2007). Social interactions and the demand for sport: An economic analysis. *Contemporary Economic Policy,* 25(4), 518-537.

Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association, 78*(383), 605-610.

Duan, N., Manning, W.G., Morris, C.N., & Newhouse, J.P. (1984). Choosing between the sample-selection and the multi-part model. *Journal of Business and Economic Statistics, 2*(3), 283-289.

Eberth, B., & Smith, M.D. (2010). Modelling the participation decision and duration of sporting activity in Scotland. *Economic Modelling, 27*(4), 822-834.

García, J., Lera-López, F., & Suárez, M.J. (2011). Estimation of a structural model of the determinants of the time spent on physical activity and sport: evidence for Spain. *Journal of Sports Economics, 12*(5), 515-537.

García, J., & Suárez, M.J. (2020). Organised and non-organised physical activity among children in Spain: The role of school-related variables. *European Sport Management Quarterly*, 20(2), 171-188.

Garrues, M.A., Lera-López, F., & Suárez, M.J. (2017). The correlates of physical activity among the population aged 50-70 years. *Retos*, *31*, 181-187.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153-161.

Hovemann, G., & Wicker, P. (2009). Determinants of sport participation in the European Union. *European Journal of Sport and Society, 6*(1), 51-59.

Humphreys, B.R., & Ruseski, J.E. (2009). *The economics of participation and time spent in physical activity*. Working Paper No. 2009-09, Department of Economics, University of Alberta.

Humphreys, B.R., & Ruseski, J.E. (2011). An economic analysis of participation and time spent in physical activity. *The B.E. Journal of Economic Analysis & Policy*, *11*(1), article 47.

Humphreys, B.R., & Ruseski, J.E. (2015). The economic choice of participation and time spent in physical activity and sport in Canada. *International Journal of Sport Finance, 10*(2), 138-159.

Jones, A.M. (2000). Health econometrics. In A. J. Culyer, & J. P. Newhouse (Eds.), *Handbook of Health Economics*, volume 1a (chapter 6, pp. 265-344). North Holland.

Kokolakakis, T., Lera-López, F., & Panagouleas, T. (2012). Analysis of the determinants of sports participation in Spain and England. *Applied Economics, 44*(19-21), 2785-2798.

Lera-López, F., & Rapún-Gárate, M. (2007). The demand for sport: Sport consumption and participation models. *Journal of Sport Management, 21*(1), 103-122.

Madden, D. (2008). Sample selection versus two-part models revisited: The case of female smoking and drinking. *Journal of Health Economics, 27*(2), 300-307.

McKelvey, R.D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*(1), 103-120.

Meltzer, D.O., & Jena, A.B. (2010). The economics of intense exercise. *Journal of Health Economics, 29*(3), 347-352.

Muñiz, C., & Downward, P. (2019). The outcomes related to sport and physical activity: A better understanding health, social, labour and academic impacts. In J. García (Ed.), *Sports (and) Economics* (pp. 393-423), FUNCAS, Social and Economic Studies, 7.

Muñiz, C., Rodríguez, P., & Suárez, M.J. (2014). Sports and cultural habits by gender: an application using count-data models. *Economic Modelling*, 36, 288-297.

Rhodes, R.E., Janssen, I., Bredin, S.S.D., Warburton, D.E.R., & Bauman, A. (2017). Physical activity: Health impact, prevalence, correlates and interventions. *Psychology & Health, 32*(8), 942-975.

Ruseski, J.E., Humphreys, B.R., Hallmann, K., & Breuer, C. (2011). Family structure, time constraints, and sport participation. *European Review of Aging and Physical Activity, 8*(2), 57-66.

Thibaut, E., Eakins, J., Vos, S., & Scheerder, J. (2017). Time and money expenditure in sports participation: The role of income in consuming the most practiced sports activities in Flanders. *Sport Management Review, 20*(5), 455-467.

WHO (2010). *Global recommendations on physical activity for health.* Retrieved from: https://www.who.int/dietphysicalactivity/publications/9789241599979/en/.

Wicker, P., Downward, P., & Lera-López, F. (2017). Does regional disadvantage affect health-related sport and physical activity level? A multi-level analysis of individual behaviour. *European Journal of Sport Science, 17*(10), 1350-1359.

Yen, S.T., & Jones, A.M. (1996). Individual cigarette consumption and addiction: a flexible limited dependent variable approach. *Health Economics, 5*(2), 105-117.