doi:10.1111/pirs.12118

Modelling regional lottery sales: Methodological issues and a case study from Spain*

Rose Baker[1], **David Forrest**[1], **Levi Perez**[2]

[1] Salford Business School, University of Salford, Salford M5 4WT, UK (e-mail: R.D.Baker@salford.ac.uk, D.K.Forrest@salford.ac.uk)

[2] Facultad de Comercio, Turismo y Ciencias Sociales Jovellanos, Universidad de Oviedo. Luis Moya Blanco, 261. 33203 Gijón, Asturias Spain (e-mail: lperez@uniovi.es)

Received: 24 July 2013 / Accepted: 26 February 2014

Abstract. Sales are modelled for 50 Spanish provinces over 522 draws of a high prize lotto game. A crossed random effects model allows for random shocks specific to particular draws and to particular provinces. The paper explores how demographic and socio-economic factors influence sales volumes, with particular focus on the relationship between sales and real *per capita* income at different jackpot sizes. When jackpots are low, sales are shown to rise with the level of provincial incomes. But this relationship disappears or is reversed when rollovers raise the jackpot because sales in poorer provinces are markedly more responsive to jackpot size.

JEL classification: C51, H27, L83, R10

Key words: Lotto demand, panel data, crossed random effects

1 Introduction

A substantial literature, surveyed by Grote and Matheson (2011), reports modelling of sales of state lottery games either at the national level or at the level of the individual household. National studies tend to focus on the variation of sales from draw to draw in response to variations in the size of the jackpot pool, usually to assess whether net revenue could be increased by changing the pay-out rate or the prize structure of the game. Household studies typically relate propensity to purchase to socio-demographic characteristics, with special reference to the incidence of the heavy taxes embedded in lottery games as between more and less affluent households, reflecting concern that taxes may be regressive. Both types of study are numerous, which is unsurprising given the importance of lottery revenue to governments. In 2010, worldwide sales of lotto games were US$245b (www.elottery.com/markets) and typically only 40–60 per cent of revenue is paid out in prizes.

* The authors acknowledge financial support and provision of data from Sociedad Estatal Loterías y Apuestas del Estado.

With notable exceptions, such as Garrett and Coughlin (2009), Coughlin and Garrett (2009) and Ghent and Grant (2010), who used county-level data for particular US states, and Oster (2004) and Combs et al. (2008), who focused on sales in smaller (zip code) areas in Connecticut and Minnesota respectively, there is a relative paucity of studies exploring determinants of sales across spatial units within a single lottery jurisdiction – despite the popularity of modelling lottery demand at the national and household levels. Further, studies which do consider spatial variation of sales do so with quite small geographical units of analysis. By contrast, the present paper seeks to explore the characteristics of regions which result in differences in baseline sales. Further, it models how the sensitivity of sales to the size of jackpot varies according to these characteristics, asking in particular whether elevated sales when the jackpot becomes very large are derived more from richer or from poorer regions.

What is the relevance of modelling sales at the regional level? First, it potentially provides guidance to operators in understanding their markets and pursuing the objective they are typically set, to maximize revenue for the state. Second, there are tensions in many countries over whether the pattern of national government expenditure and revenue is equitable between regions. In all jurisdictions, lottery tickets are subject to a heavy explicit or implicit tax which ultimately accrues to national governments or to the organizations (for example, sports governing bodies) it nominates to be supported by lottery proceeds. It is of interest whether this revenue might, for example, be drawn disproportionately from disadvantaged regions (before, perhaps, being spent on facilities, such as opera houses, in more advantaged regions). Third, it is known that lottery agencies enjoy particularly high sales when a large jackpot is available for a particular draw. Do these additional sales derive most from the poorest or from the richest regions? What does the answer imply about motives for buying lottery tickets? Are big jackpots or small jackpots more associated with regressivity?

Kitchen and Powells (1991) appears to be the only prior paper on lotto sales which disaggregates to the regional level. It examined determinants of expenditure in six regions of Canada. However, it used household data and simply estimated a tobit equation, for each region separately, to relate expenditure to household characteristics. It may be noted that in every region, expenditure was positively related to income and negatively related to a categorical variable signifying that the head of household had a degree. A categorical variable for 'urban' was positive and significant in three cases. These relationships from household studies suggest that measures of income, education and urbanization should also be tested for inclusion in spatially aggregated sales equations.

Our purpose is the general one of exploring how regional sales equations should be estimated for state lotteries. But, of course, we work in a particular context. This is provided from Spain.

2 The setting

We were able to observe sales of a national lottery game for each of the 52 provinces of Spain for each of 522 weekly draws between the start of 1998 and the end of 2007.

The game is *El Gordo de la Primitiva*, a long-odds, high jackpot game played once a week. As with other lotto games, players choose a set of numbers and any players whose set exactly matches the randomly-drawn winning combination share the jackpot pool; secondary pools pay prizes for near misses. If no player wins the jackpot pool, the amount in it is carried forward to the next draw and this will continue until a draw when at least one player wins. Such 'rollovers', if repeated for several weeks, can result in dramatically high jackpot prize levels, as high as €25.7m in our data set. Rollovers produce considerable variation in value for money from draw to draw and thereby facilitate identification of the relationship between sales and any measure

of value for money that the researcher employs. Forrest et al. (2002) demonstrated a closer relationship between sales of UK lotto and jackpot size than between sales and the expected value of holding a ticket and García and Rodríguez (2007) confirmed this result for *La Quiniela*, the national football pools game in Spain. Here, therefore, we use jackpot size as our relevant explanatory variable. Although this is observed only *ex post* (since the jackpot pool includes a fixed fraction of that draw's sales), its amount should be closely anticipated by players and potential players on the basis of the information, provided by the lottery agency, on the amount of rollover funds in the pot for the current draw.

It should be noted that, as in other European countries, Spain offers a portfolio of games on different days of the week and with different prize structures. Among the national games (i.e., apart from the transnational *EuroMillones*), *El Gordo de la Primitiva* typically offers the highest jackpot. However, Perez and Forrest (2011) found little cross-elasticity between the various games, so that a high jackpot in one did not cannibalize sales of the others in the same or following week.

The data made available to us by the operator extend over the ten years from 1998 to 2007. Throughout this period, the entry fee to *El Gordo de la Primitiva* was the same, €1.50 (prior to the adoption of the euro, tickets were priced in pesetas but the euro equivalent was so close to €1.50, in fact 1.503, that we feel justified in ignoring this change in what follows). The proportion of sales revenue earmarked to be returned in prizes (0.55) was also constant through the period. However, the format of the game was changed in February 2005 (despite there having been no tendency for sales to decline over time, as occurred in many countries). Previously, players had selected six numbers from the set 1–49, the familiar 6/49 game offered as the core lotto game in many jurisdictions. Under the new design, players chose from two matrices, selecting five numbers from the set 1–54 in matrix 1 and one number from the set 1–9 in matrix 2. This change lengthened the odds of a perfect match considerably, from about 14m to 1 to about 32m to 1. Other rule changes included a minimum guaranteed jackpot of €5m (to make draws with no rollover money more attractive) and extra prize tiers for near misses. These changes made for longer sequences of draws without a winner and altered the pattern of demand from draw to draw. The break in behaviour is clearly evident in the time series of national sales (Figure 1, where a vertical broken line shows when the format changed).

The most obvious change in behaviour is that week to week sales showed much less variation following the design change. The draws where sales were weakest after February 2005 attracted far higher sales than the corresponding 'low' weeks before: this effect was to be

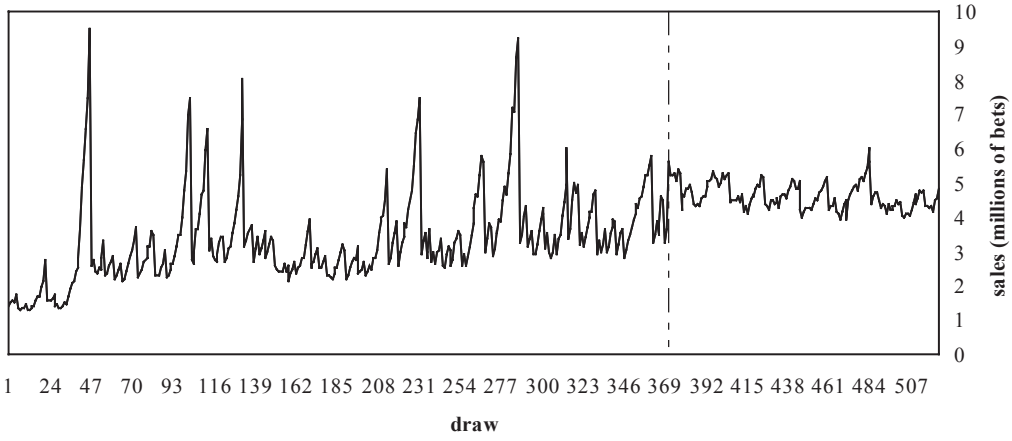

Fig. 1. Draw by draw national sales (number of tickets sold) of *El Gordo de la Primitiva*

expected because of the innovation that a minimum jackpot of €5m was now guaranteed whereas, earlier, draws without rollover money had offered much lower levels of jackpot. The peaks in the time series of sales also became less extreme. This may be related to the fact that the game became harder to win and draw cycles (the sequence of draws without a winner) therefore tended to be longer. Some players would then likely postpone participation or the purchase of extra tickets because there was a fair chance that a draw with a given jackpot would produce no winner and an even higher jackpot would be available the following week.

Below, sales are modelled at regional rather than national level and some indication therefore needs to be provided of how these regional units are derived. Spain is divided into 17 autonomous communities. Each contains provinces (sometimes only one) and it is the 50 provinces across Spain which define our spatial unit. In addition, there are two autonomous cities, Ceuta and Melilla, which are small Spanish enclaves in North Africa and which we treat as provinces. This gives the final total of 52 cross sectional units, 'provinces'.

The provinces are heterogeneous in population, geographical extent and degree of urbanization. For example, population in the final year of our data period ranged from less than 70,000 to more than 6m. But it would be fair to say that the archetypical province contains a central city in its interior, together with satellite centres and a rural hinterland. Cross-border sales are likely to be unimportant given that most people live in the interior of a province and this suggests that, when we observe sales in a province, we also observe demand by residents of that province. This should sharpen the relationship between sales and demographic and socio-economic characteristics compared with estimation over smaller spatial units.

Draw by draw sales data at the province level and jackpot data for all draws were supplied to us by the lottery agency, Sociedad Estatal Loterías y Apuestas del Estado (SELAE). SELAE sells tickets to the game through a network of outlets located throughout Spain (including Ceuta and Melilla). Demographic information on provinces was from INE, the official statistical agency for Spain. Data on education level by province were collected from IVIE (the Valencian Institute of Economic Research) and figures on household disposable income from the Fundación de las Cajas de Ahorros (the Foundation of Spanish Savings Banks).

3 Developing the statistical model

3.1 Exploratory analysis

We have a balanced panel of 52 provinces observed at each of 522 time points. We seek to account for the sales variable q_{it}, where q is the (*per capita*) number of tickets sold and the subscripts i and t index provinces ($I = 1, 2, \ldots, 52$) and draw numbers ($t = 1, 2, \ldots, 522$) respectively. Here and subsequently, all references to sales refer to sales *per capita* where the population is defined by that aged 18 years or over, the legal age for gambling in Spain.

Pooling of data, and then applying ordinary least squares (OLS) regression, is likely to be problematic given that cross-sectional correlation can be an important feature of spatial panel data. Baltagi (2008) discusses models appropriate for use in modelling spatial panels, depending on the importance of cross-sectional correlation.

In preliminary analysis to assess the importance of cross-sectional correlation, simple sales equations were estimated for each province separately, with log sales at draw t specified to depend on a set of 'lottery' variables (lagged log sales, log jackpot, categorical variable for the new format of game and trend (identified separately for the sub-periods with the old and new game formats)). Specifying both sales and jackpot in terms of their natural logs permits coefficients to be interpreted as the elasticity of sales with respect to jackpot. For now, variables representing demographic and socio-economic variables in each province, which change only slowly, were

omitted and their effects will have been reflected in the respective constant terms. That coefficient estimates on log jackpot proved different between provinces suggested that they indeed depend on socio-economic factors, providing provisional support for our suspicion that, for example, better and worse off provinces might respond differently to the stimulus of a high prize.

After the regression equations had been estimated for each province, we inspected the residuals at each time point and examined the correlation in the residuals as between each pair of provinces. Let r_{ab} be the correlation coefficient between the T residuals for province a and the T residuals for province b (here $T = 521$ because, although there were 522 time points, a lagged dependent variable was included in the regression equation and so the first set of observations, from $T = 1$, was not used). The mean value of r (which was always positive) across all 52×51 pairs of provinces was 0.771 (standard deviation 0.139). Now define $r_{ab}^{(1)}$ where we are measuring this time the correlation between the residual for province a for draw t and the residual for province b at draw $t\text{-}1$ (observations of the residuals for the first draw of the 521 were not used). We took the mean of $r_{ab}^{(1)}$ and $r_{ba}^{(1)}$ and averaged across the corresponding statistic for all 52×51 pairs of provinces. The mean pairwise correlation was then 0.386 (standard deviation 0.095), much lower than before. Together, the results imply that there are unobserved shocks specific to particular draws which affect total aggregate sales. This is cross-sectional correlation which needs to be taken into account in the selection of the statistical model (it rules out applying OLS to pooled data). Examples of shocks of this nature might be a national advertising campaign for the lottery or a major television news event which stopped players across the country from going to the lottery shop.

Another potential layer of complexity is that there may be unobserved shocks which affect sales in particular draws just in one part of Spain, for example, storms might from time to time keep lottery players at home but only in the North West. If such shocks are important, residuals for particular pairs of provinces, for example, adjacent provinces, could be highly correlated. For all 52×51 pairs of provinces, we plotted r_{ab} against the distance between a and b. But in fact there was no relationship between fit residuals between provinces and distance. This simplified modelling since provinces otherwise would have had to be thought of as being embedded within wider regions making up the national market or else a spatial error regression model (see Baltagi 2008) employed.

3.2 Choice of covariates

Log sales in each province in each lottery draw are modelled as a function of the characteristics of that draw (as outlined above, including the log of the size of jackpot) and of the demographic and socio-economic characteristics of the province. The variables for the lottery change from draw to draw, that is weekly, but the variables describing the province (e.g., gross disposable household income, defined *per capita*) are typically issued only annually. Weekly values for all variables in this category were interpolated from annual values using cubic splines. This is a quite flexible approach to the problem and is widely employed in regression analysis (see Harrell 2001). Where regressors were measured in money (jackpot size and *per capita* household disposable income), figures were deflated by the consumer price index to make them 'real' rather than 'nominal'. Consequently, all are measured in euros with the purchasing power of December, 2007 (the final month in our data period).

This gave the following covariates for inclusion in the model: lottery variables:

(*i*) log *per capita* sales lagged one draw;
(*ii*) log real size of jackpot (and its square, to allow jackpot elasticity of sales to vary with jackpot);

(*iii*) categorical variable for new game format (= 1 from February 2005);
(*iv*) draw number (trend) if old game format in place (= 0 from February 2005); and
(*v*) draw number (trend) measured from introduction of new game format (= 0 before February, 2005) .

Province variables:

(*vi*) log real gross disposable household income per adult (and its square, to allow income elasticity of sales to vary with income);
(*vii*) mean years of completed education;
(*viii*) log proportion of total population under 18 years;
(*ix*) log proportion of total population 65 years or over;
(*x*) log of population density (population over 18 divided by province area in sq. km.); and
(*xi*) categorical variable = 1 if province is in Catalonia.

Interaction terms:

(*xii*) log real gross disposable household income per adult times log real size of jackpot;
(*xiii*) log mean years of completed education in the province times log real size of jackpot; and
(*xiv*) log population density times log real size of jackpot.

In the case of squared and interaction terms, variables were defined to be centred on their mean values in the most recent year in the data set. This permits coefficient estimates on log income and log jackpot to be interpreted as elasticity estimates for the case where variables take on the mean values observed in 2007.

Among the lottery variables, the inclusion of a lagged dependent variable allows for habit formation and is a typical feature in time series modelling of lotto sales. A significant coefficient here should not be assumed to indicate 'addiction' as it may simply be that buying a ticket at a particular time one week serves as a 'reminder' when that time arrives again the following week. A significant coefficient estimate might also, in the case of Spain, reflect reinvestment, since 10 per cent of tickets each week are awarded a refund of the purchase price through a random process related to ticket serial numbers. These refunds are collected at lottery sales outlets.

Variables (*iv*) and (*v*) are trend terms to control for the possibility that sales drift down over time following the introduction or relaunch of a game. The notion that players become bored and disillusioned with time is supported by several time-series studies of demand for lotto games in various jurisdictions.

Among province variables, measures for income, education and population density (a proxy for urbanization) are suggested for inclusion by results from household level demand studies. Variables (*viii*) and (*ix*) will reflect information on family structure. Variable (*viii*), the proportion of minors in the whole population, may be relevant because any given level of *per capita* income is likely to be less adequate where there are more children to be supported. There may also be a direct negative effect from variable (*ix*) since retired persons have often been noted (e.g., in Forrest and Gulley, 2009) to have a below average propensity to play the lottery. The categorical variable for Catalonia (where four provinces are situated) is included because the Catalan government operates its own lottery with a standard suite of games and this represents competition for the national games offered by SELAE. Interaction terms are included to test whether, for example, the sensitivity of sales to size of jackpot varies with province characteristics.

3.3 A crossed random effects model

The model suggested by the exploratory analysis was a crossed random effects model:

$$q_{it} = \Sigma\beta X_{it} + \eta_i + \xi_t + \varepsilon_{it}. \quad (1)$$

Here, subscripts i and t refer to the province and draw number respectively. The dependent variable is log *per capita* sales (in province i in draw t). The covariates X refer to the 14 explanatory variables listed above (and an intercept term). The three error terms, η, ξ and ε are independent normally distributed random variables, each with mean of zero. The random effect η_i allows for features of the ith province that the covariates cannot capture, the random effect ξ_t allows for features of the tth draw that the covariates cannot capture, and ε_{it} is the usual 'pure' error term.

Panel data analysis poses many statistical problems which use of this model, together with extensions noted below, seeks to address. These include heterogeneity (non-normality of fit residuals) and correlations between successive fit residuals and/ or between fit residuals from different provinces.

Heterogeneity was dealt with by applying a Box-Cox transformation to the dependent variable, which generalizes the model by allowing the residuals to become more Gaussian. The transformation parameter λ was then estimated by maximum likelihood along with the other parameters. This step, involving computation using a purpose written program, improved model fit substantially but did not change predictions by much. Correlation between successive fit residuals was dealt with by including lagged sales in the model. Correlation between fit residuals from different provinces was ruled out by the exploratory analysis which found no relationship between r_{ab} and distance between a and b.

Endogeneity could also, in principle, have been a problem. Sales of tickets depend on jackpot but also contribute to jackpot because a percentage of sales revenue is paid into the jackpot pool. In modelling of national lottery sales, this source of endogeneity bias in estimation is typically addressed with an instrumental variables model, with jackpot size (or expected value, depending on choice of explanatory variable) instrumented on size of rollover. We judged this probably unnecessary here. The feedback from sales to jackpot will usually be small at the province level because a single province typically contributes only a low proportion of the jackpot pool. Over the whole period covered by the data set, only Madrid (16.2%) and Barcelona (9.2%) among the 52 provinces accounted individually for as much as 6 per cent of aggregate sales. However, as a precaution against contamination of our coefficient estimates, we present 'lead' results based on excluding observations from these two large provinces, which purchased, between them, a quarter of all tickets. For transparency, we report also results based on all 52 provinces (including Madrid and Barcelona). In practice, results were in fact very similar, illustrating that it is unnecessary in modelling regional lottery sales to increase complexity further by correcting for endogeneity (providing of course that, as here, the number of regional units defined is sufficiently large).

Discussions of panel data estimation also raise the issue of fixed versus random-effects models. Use of a fixed effects model here was not feasible. Giving each province its own fixed effect would make it impossible to study the dependence of sales on demographic and socio-economic variables which differ across provinces but do not vary much with time. The same applies to giving each lottery draw its own fixed effect since then we could not study the dependence of sales on jackpot size.

A random effects model, though a necessary choice for us, is in fact a more ambitious model than a fixed effects model, because it models sales entirely in terms of explanatory variables. On the other hand, there is a general problem that the explanatory variables may correlate with the

random effect, a source, albeit one hard to motivate, of possible endogeneity of the various regressors.

Greene (2011) discusses the use of the Hausman specification test and also the approach of Mundlak (1987). We followed Mundlak's recommendation to generalize the model by adding extra parameters, with the test of whether these parameters are non-zero equivalent to a Hausman test. We computed mean real *per capita* household disposable income over the whole data period for each province and used this as an additional covariate in the regression. The resulting model (which cannot be used for prediction as it specifies sales as dependent on future data) enables a test for endogeneity. Here the additional regressor proved not to be significant, showing that that there is no evidence that the random effects model is inadequate by virtue of endogeneity.

4 Results

4.1 General considerations

Some key summary statistics are displayed in Table 1 and the results themselves in Table 2. In Table 2, the first column presents our lead results which include all covariates (and are based on excluding the two largest provinces from the sample as a precaution against endogeneity). The second column also shows the results from a crossed random effects model but with the difference that only lottery and income variables are included in the specification (i.e., except for income, variables describing the characteristics of provinces are excluded). The third column reports results from the full model when observations for Madrid and Barcelona were reinstated. The fourth column displays OLS results.

Table 1. Key summary statistics

	Mean	sd	Min	Max
Whole period				
Tickets sold per eligible adult	0.103	0.040	0.021	0.761
Log tickets sold per eligible adult	–2.342	0.387	–3.881	–0.274
Real jackpot (millions of euros)	5.731	5.058	0.385	25.845
Log real jackpot	15.127	1.004	12.862	17.068
Real gross household disposable income *per capita* (euros per year)	17,735	2,937	11,989	27,793
Log real gross household disposable income *per capita*	9.770	0.160	9.392	10.232
Proportion of population under 18	0.177	0.031	0.118	0.282
Proportion of population over 65	0.188	0.045	0.101	0.285
Mean years of education	9.407	0.658	7.940	11.260
Population density (adults per sq. km.)	203.59	624.86	7.46	3,650.95
Final year only				
Tickets sold per eligible adult	0.121	0.022	0.066	0.213
Log tickets sold per eligible adult	–2.130	0.190	–2.715	–1.546
Real jackpot (millions of euros)	9.783	5.496	5.000	25.845
Log real jackpot	15.972	0.473	15.425	17.068
Real gross household disposable *income per capita* (euros per year)	18,971	3,096	13,520	27,570
Log real gross household disposable income *per capita*	9.838	0.157	9.512	10.224
Proportion of population under 18	0.170	0.030	0.118	0.264
Proportion of population over 65	0.186	0.043	0.106	0.285
Mean years of education	9.840	0.568	8.320	11.260

Notes: Summary statistics refer to observations on the fifty provinces included in the principal regression results (i.e. observations on the two largest provinces were not used in the calculation of the summary statistics).

Table 2. Estimation results (dependent variable is log sales)

model	(1) CRE	(2) CRE	(3) CRE	(4) OLS
two largest provinces	excluded	excluded	included	excluded
log *per capita* sales lagged one draw	0.834***	0.840***	0.836***	0.711***
	(0.003)	(0.003)	(0.003)	(0.003)
log real size jackpot	0.037**	0.034**	0.037**	0.080***
	(0.005)	(0.014)	(0.014)	(0.002)
square of log real size jackpot	–0.035***	–0.035***	–0.022***	–0.023***
	(0.005)	(0.005)	(0.003)	(0.001)
new game format	–0.111***	–0.109***	–0.112***	–0.129***
	(0.024)	(0.024)	(0.024)	(0.004)
Trend × old game format	0.000	0.000	0.000	0.0001***
	(0.000)	(0.000)	(0.000)	(0.000)
Trend × new game format	–0.000	0.000	0.000	–0.0001***
	(0.000)	(0.000)	(0.000)	(0.000)
Log real *per capita* disposable income	0.015	–0.017	0.015	–0.012**
	(0.011)	(0.011)	(0.011)	(0.009)
Square of log real *per capita* disposable income	0.033	–0.055**	0.010	–0.117***
	(0.025)	(0.022)	(0.025)	(0.002)
Mean years of completed education	0.007***		0.008***	–0.012***
	(0.002)		(0.002)	(0.003)
Log population density	–0.003		–0.004	0.009***
	(0.004)		(0.003)	(0.001)
Log proportion of population aged 65+	–0.093***		–0.099***	–0.159***
	(0.017)		(0.016)	(0.009)
Log proportion of total population aged <18	–0.147***		–0.152***	–0.275***
	(0.002)		(0.020)	(0.011)
Catalonia	–0.033**		–0.037***	–0.040***
	(0.015)		(0.013)	(0.004)
Income variable × jackpot size variable	–0.023***	–0.022***	–0.022***	–0.015**
	(0.003)	(0.002)	(0.003)	(0.007)
Education variable × jackpot size variable	0.002**		0.002***	–0.012***
	(0.001)		(0.001)	(0.002)
Population density × jackpot size variable	–0.0006**		–0.0003	0.0008
	(0.0003)		(0.0003)	(0.0006)
Intercept	–2.115***	–2.133***	–2.115***	–2.114***
	(0.019)	(0.020)	(0.019)	(0.004)
Residual standard deviation	0.052	0.052	0.052	
Province sd ratio	0.471	0.553	0.460	
Temporal sd ratio	2.100	2.105	2.089	
Number of cross section units	50	50	52	
Number of temporal units	521	521	521	
Pooled sample size				26,050

Notes: CRE refers to crossed-random effects model, OLS to ordinary least squares. Standard errors are shown in parentheses. Coefficient estimates and standard errors are shown to three decimal places (or four decimal places where the fourth was the first significant figure). *** and ** indicate statistical significance at the 1% and 5% levels respectively.

While the OLS model was successful in terms of goodness of fit achieved (R^2 was 0.904), the constraint of not distinguishing random effects due to draw and province is shown to be potentially damaging. On adding the two random effects (i.e., moving to the crossed random effects model) the chi-squared on two degrees of freedom is 34,370. This is convincing evidence that the two random effects 'belong' in the model, and their standard deviations are roughly half

that of the residual error for province and twice it for draw, suggest that OLS over or underestimates sales for particular draws across all provinces. Omitting the random effects has practically important consequences since the coefficient estimates on some variables (e.g., education) change sign between columns 1 and 4 while others (e.g., population density and the two trend terms) change from highly significant to insignificant. But significance levels will tend to be inflated if panel data are estimated by OLS (Goldstein 2010); here, statistical significance would be evaluated as if there were 50×521 independent observations rather than only 521 observations on each of 50 cross sectional units. Magnitudes of other coefficient estimates, such as on lagged sales and real jackpot, change substantially. All this illustrates that it is of practical importance when modelling demand across spatial units to take care to select the appropriate statistical model. It is emphatically not the case that settling for a naïve model will suffice because only the details of the results will be affected.

Our commentary section relates to findings from models described in the first two columns. The two models differ to the extent that the full set of covariates appears in model 1 whereas income is the sole descriptor of a province employed in model 2. The reason for having two models is that different questions may be asked of them. The first, for example, allows the modeller to predict effects on sales of a 10 per cent fall in the level of real household incomes, such as could occur in the age of austerity. If incomes fall because of a recession, other influences should be held constant because the modeller is contemplating a situation where variables such as those for education and population would not change (at least for a long time). The necessary *ceteris paribus* conditions are imposed by the full specification of model 1. Essentially, the modeller is estimating what in household level studies would be called the Engel curve: he is showing how purchases of the product change as income changes, *ceteris paribus*.

On the other hand, the regressivity issue relates to the simpler question of whether poorer or richer provinces buy more lottery tickets and therefore contribute more implicit tax. Here, it would not matter to the questioner why provinces have different levels of income. For example, whether poor regions are poor because their populations have low education levels or for altogether different reasons should not change the answer to the question posed. Yet, if education remains in the model, its coefficient may absorb much of any relationship between sales and income. Education and other variables must therefore be omitted from the model if the regressivity question is to be correctly answered: in this case, the estimate of the marginal effect of income on sales should not be conditioned on education and other variables.

In all sets of results, the coefficient estimate on the lagged dependent variable is large and significant. This is important for interpretation of the coefficient estimates on the focus variables. Where the right hand side variable is logged, the coefficient estimate is a direct measure of the short-run elasticity of sales with respect to that variable. Long-run elasticity is obtained by multiplying the coefficient estimate by $1/(1\text{-}b)$, where b is short-run elasticity. In the case of our lead results, this implies that the estimate of long-run elasticity is approximately 6.04 times as large as the corresponding coefficient estimate for the variable of interest.

Whether short or long-run elasticity is the appropriate estimate depends again on the question put. For example, the short-run elasticity with respect to real jackpot enables the modeller to predict how sales will respond to an increase in jackpot caused by a rollover. In this case, the change in jackpot would relate only to the current draw. But, if the operator were to consider raising the proportion of revenue paid out in prizes and putting all of this increase into the jackpot pool, it would be long-run elasticity which would predict sales in the new equilibrium because the boost to the jackpot would be permanent. In the case of income, the consequences of a long-run decline in real incomes would similarly be revealed by long-run rather than short-run elasticity as would be differences in sales levels between provinces with different degrees of long-run economic prosperity. That long-run elasticity is several times larger than

short-run elasticity should therefore be borne in mind when interpreting any relationship between sales and income revealed by coefficient estimates on real income.

4.2 Commentary

Our main focus is on the relationship between sales and income and how this varies according to the size of jackpot. However, we first offer brief comments on results on the control variables included in the full model as displayed in Table 2, column 1.

Among the lottery control variables, the change in game format is represented by shift and slope dummy variables. Trend terms are insignificant but the shift dummy, *new format*, proves to be negative, indicating that significantly lower sales were associated with any given jackpot once the new format was in place. This does not imply that annual sales suffered from the revision of the structure of the game since making the game harder to win was designed to produce more rollovers and greater frequency of high jackpots, making for more weeks with high sales volumes (this is what, in fact, happened: comparing the twelve months preceding and following the change in design, aggregate national sales increased by 23.6%). It does imply that a proportion of players found the new rules off-putting, for example, it is more effort to choose numbers from two matrices rather than one. Some players experiencing 'entrapment' may also have taken the opportunity to exit the market. Entrapment (Wolfsen and Briggs 2002) refers to the situation of players who always played the same numbers under the old format and were afraid to stop for fear of the regret they would feel if those numbers subsequently came up as a winning combination. A new format, with a different specification of numbers to be selected, ends that possibility.

The categorical variable for Catalonia suggests some loss of sales from competition with the regional lottery. Combining the coefficient estimate with that on lagged sales to derive an estimate of the long-run effect, *per capita* sales in provinces located in Catalonia were about one-fifth lower than might have been predicted from the other covariates.

We noted in the Introduction that the degree of urbanization should be taken into account in modelling sales and this possibility is investigated here by inclusion of the variable population density. It might plausibly be anticipated that sales would be relatively low in less densely populated areas because it would be harder to supply a well scattered population. But the results fail to support this hypothesis, suggesting perhaps that the operator has been successful in Spain in providing a retail network adequate for ensuring that potential purchasers are not deterred by the difficulty of accessing points of sale. Curiously, there is an unexpected negative sign on the interaction term including population density, such that sales in more urban provinces appear to respond less to increases in jackpot compared with more rural provinces.

Results on the demographic variables return signs that might have been anticipated. The proportion of the eligible population aged over 65 is a strongly significant negative determinant of sales, consistent with findings in household studies (e.g., Forrest and Gulley 2009) that the oldest age group displays relatively low engagement with lottery games. The proportion of the total population below 18 is also a significant negative predictor. The model controls for real disposable income per adult, so it is unsurprising that lottery purchase falls with the number of children since the extra cost of children will reduce the scope for discretionary spending, such as that on lottery tickets, to be funded from any given level of income.

Education variables have had mixed results in prior literature. Our education variable is mean years of education in a province, similar to Forrest and Gulley (2009) who had age at which the head of household left full-time schooling. Forrest and Gulley found from UK household expenditure data that a higher level of schooling was associated with a fall in both the probability of participating in lotto and the level of play conditional on participation. Kitchen

and Powells (1991) also used household level data and analysed lotto purchases in each of six regions of Canada; in every case an indicator variable for 'graduate' was negative and significant in tobit estimation.

But in spatial studies of demand at the level of US counties or zip codes, while Giacopassi et al. (2006) found the negative relationship between sales and the proportion of graduates in a county that might have been expected from household studies, Price and Novak (1999) reported the opposite for lotto in Texas; and Oster (2004) found that the proportion of graduates in a zip code played no role in predicting sales of Powerball. In our present results, for spatial units that aggregate many more households than in the American research, we find that education is a strongly positive predictor of sales (further, the significant positive coefficient estimate on the interaction term between education and real jackpot indicates that this relationship strengthens in higher jackpot draws).

The contrasts in conclusions from the various studies may reflect their different levels of aggregation. On the other hand, there is likely also to be a difference arising from how spartan the specification is. Using household expenditure surveys, Kitchen and Powells (1991) and Forrest and Gulley (2009) had very large sample sizes and could employ a very rich set of controls. For example, Forrest and Gulley included not only income but also occupational status (defined by categories from unskilled to professional). They found that, controlling for income and education, the group with the most enthusiasm for the lottery was 'intermediate', where the head of household had a semi-skilled manual or junior white collar job. In such as Oster's (2004) study and in ours, the influences of occupational status and education may simply be conflated in the single variable measuring schooling and this is another possible explanation of any inconsistencies. In the final year of our data, mean years of education ranges from 7.9 to 11.2 years, roughly equivalent to incomplete versus complete secondary education, and it would not be unreasonable to suppose that the more educated provinces had higher proportions of the labour force in medium compared with low status occupations. Of course this is speculative. In a spartan specification, mean years of schooling will reflect many aspects of social structure; but its importance in the results means that, whatever the cause, the variable is serving as an effective control to help isolate the relationships between sales, jackpot and income.

Regarding the influence of jackpot size, the coefficient estimates on the level and squared terms are highly significant and signed positive and negative respectively: a given proportionate increase in the size of the jackpot raises sales but at a decreasing rate. The real income terms, by contrast, are insignificantly different from zero.

But both these results are of limited interest since inspection of the relevant coefficient estimates reveals that the interaction between the jackpot and income terms is very important. Because one has to take account of the fact that both jackpot and income are specified as quadratics and the interaction term is highly significant (which implies, e.g., that elasticity with respect to income may be zero at mean income and mean jackpot but may not necessarily be so in draws with different jackpot size or for provinces with higher or lower incomes), interpretation of the results in Table 2 is not straightforward. We therefore present diagrams to illustrate what the results mean in terms of the relationships in which we are interested.

Figure 2 plots the relationship implied by the (column 1) results between the log values of predicted per capita sales and real per capita income for four levels of jackpot between €5m (the basic guaranteed jackpot for weeks when there is no rollover money) and €20m. The range of incomes shown on the horizontal axis approximates the range of values observed in the final year of our data. Figure 3 repeats the exercise but this time for the results from the 'income only' model (column 2).

Focusing first on Figure 2, tickets for the basic draw (no rollover) are shown to be a 'normal good', that is, higher incomes are unambiguously associated with higher sales. This mirrors findings in the household level study by Forrest and Gulley (2009) where, however, there was

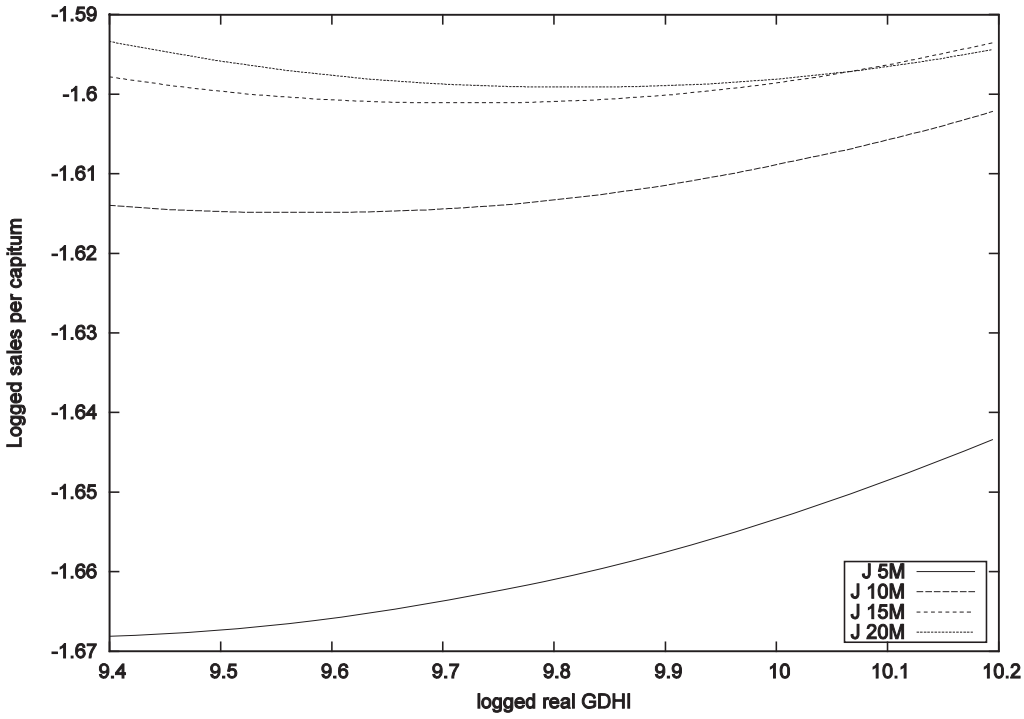

Fig. 2. Relationship between *per capita* sales *and per capita* disposable income at selected sizes of jackpot

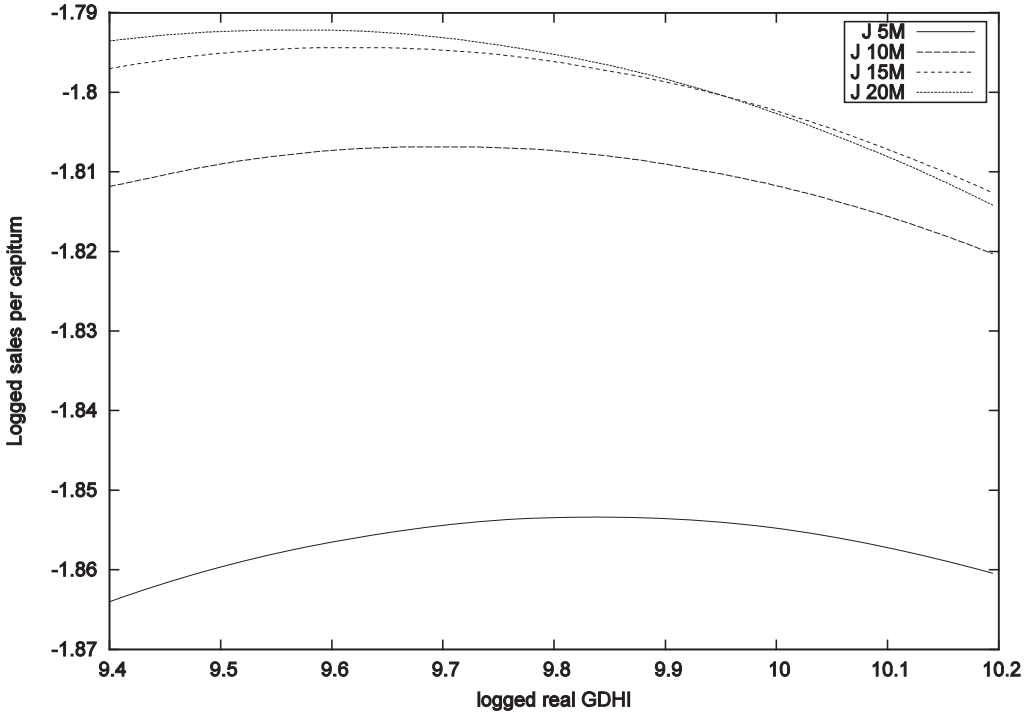

Fig. 3. Relationship between *per capita* sales and *per capita* disposable income at selected sizes of jackpot (income only model)

no allowance for size of jackpot (the UK household expenditure survey collected spending over two weeks during which there would be four lotto draws; typically, but not of course always, these would have featured more than two 'basic' draws).

It is striking from Figure 2 that the game ceases to be a normal good when higher jackpots are offered. At €10m, the relationship between sales and income is quite flat, consistent with the coefficient estimate on real income in Table 2 being zero (and elasticity also zero when all other relevant variables, including jackpot, are set equal to their final year means). Then, by the time we reach the highest size of jackpot, the relationship between sales and real income is reversed. For this size of jackpot, tickets are a mildly inferior rather than a normal good. This change arises of course because poorer provinces display higher elasticity of demand with respect to jackpot than richer provinces (though it is positive for both).

So what might the operator expect if incomes were to fall across the country? Sales in a non-rollover draw would be predicted to slide. Those in a large jackpot draw would be predicted not to fall and even to increase. It is tempting to infer that the first prediction just follows from affordability becoming less whereas, when really large amounts are on offer, the lottery has great psychological appeal to populations that are under financial pressure. But this is probably over-interpretation. An alternative explanation is that populations in low income regions find affordability an issue and spend their limited budget for lotteries exclusively in weeks when the tickets offer better value in terms of how much they might win. Falling incomes across the country might similarly make households more selective about the draws on which they focus their lottery spending.

The 'income only' model (Figure 3) yields slightly different patterns but with the same essential features that there are positive but diminishing sales returns to increases in the jackpot and that higher jackpots yield a greater proportionate response at low incomes. This time we have no other province specific variables in the model. Here we are not deriving income elasticity on a *ceteris paribus* basis but simply asking whether lottery tickets are purchased more in richer or in poorer regions. The answer for the basic draw is that sales are strongest in middle income provinces. Tickets are first a normal and then an inferior good, not untypical of entertainment goods where wider options for entertainment become available when incomes reach a high enough level. However, as jackpot grows, the relationship changes such that tickets become an unambiguously inferior good at the highest level of jackpot offered. Those who oppose state lotteries would be tempted to conclude that poor regions present concentrations of vulnerable people who may readily be tempted by large jackpots and suffer a sort of 'lotto frenzy'. For similar reasons as before, we would caution against over-interpretation.

We have focused on how sales vary between poorer and richer regions at different sizes of jackpot. The poorest regions appear to contribute proportionately more lottery tax, relative to their populations, than better-off regions when the jackpot is sizeable; but the relationship is largely reversed where only the basic jackpot (with no rollover money) is on offer. But what of the product as a whole? Over a year, there will be a mix of draws with varying levels of jackpot. Perhaps, averaged over a year, there is no relationship between sales and income level.

This, in fact, proves to be the case. Separately for each of the final five years represented in the data set (2003–2007), we regressed *per capita* sales on *per capita* income. Estimation was by OLS, with each province representing one data point. We estimated with linear, quadratic, log-linear and semi-log specifications. In no case was the coefficient estimate on the income variable statistically significant, confirming that annual sales of this particular lotto game do not vary according to the level of provincial income. Over time, better and worse-off provinces therefore contribute similar amounts to national revenue from the lottery game (though this implies that poorer provinces pay more lottery tax as a proportion of household incomes).

4.3 Conclusions from the results

We have sought to supply a template for operators seeking to develop a model of regional sales within their jurisdiction for purposes of planning or benchmarking. A key current issue for the European lottery industry is whether financial pressure on households in an age of austerity will shrink operators' markets and threaten their ability to provide a revenue stream for the state or good causes. Estimation of sales at the regional level, even with annual or quarterly data, would be valuable in planning for this possibility because it would enable past sales in parts of the country where income has always been low to provide insights into the likely situation where other regions also become relatively poor compared with their present. But estimation at the regional level using more detailed draw by draw data has been shown here to yield additional insights because the findings alert operators to the notion that austerity might change the pattern of demand across draws as jackpots vary. This might deserve a policy response since a change in game parameters would have the ability to alter the relative frequency of draws with different levels of jackpot.

In terms of public policy and welfare, there has been debate about whether high jackpots tend to increase or decrease the extent to which the tax implicit in lotto games is regressive. Employing spatial data, Ghent and Grant (2010), and Combs et al. (2008) compare different games offered by a lottery agency and report a tendency for games with higher jackpots to be less regressive. Like us, Oster (2004), examines differences according to jackpot size in a single game and finds again that high jackpot games deliver lower regressivity. But our results contradict Oster's conclusion. However, it should be noted that the spatial units in our analysis are much larger than those analysed by Oster and the other authors.

By changing the spatial unit employed in analysis (Oster used zip codes, roughly speaking neighbourhoods, whereas we use regions), we have exactly reversed her findings. In our data, spatial units with relatively low incomes exhibit a disproportionate increase in the propensity to purchase as the jackpot becomes large. These results imply that high jackpot editions of the same game deliver higher, not lower, regressivity.

Perhaps in fact it should not be expected that results should be similar when different sizes of spatial unit are employed in analysis. For example, lotteries appeal to aspiration to gain wealth and status. In a low income zip code (neighbourhood), given segregation by income in the housing market, most residents are likely to be in low status occupations, with few managerial and professional workers. In a low income Spanish province, this is likely to be true to a much lower degree because one reason for low average incomes in a whole region may be that wages are below average in any given occupation. It is plausible that aspirations and therefore purchasing behaviour with respect to lottery products may be influenced by individuals comparing themselves with others around them and this is one reason for expecting different results when there is a difference in the size of spatial unit analysed.

Coughlin and Garrett (2009) warn about a similar factor which may lead to contradictions at different levels of spatial analysis. Employing data on sales in US counties in five states, they show sharply different marginal propensities to spend on lottery products depending on the source of income (e.g., social benefits or wages and salaries). Our study, and others cited above, use measures of aggregate income in the spatial unit without regard to the break-down of categories of income. The extent to which differences in mean income between areas may be explained by differences in social benefits and wages may vary according to the size of area considered.

Caution also needs to be exercised in using analysis such as Oster's and ours to draw firm conclusions on whether lottery agencies should amend the prize structure of a given game in order to address concerns over regressivity. Both papers generate their findings by modelling sales as a function of jackpot. But all of the variation in jackpot is traced out in draw cycles

which start with a basic jackpot and proceed to higher and higher jackpots until a winner (or winners) is found. This means that variation in sales will partly be attributable to inter-temporal substitution where households' spending is withheld from the basic draw and added to their budget for when more attractive jackpots are on offer. This effect makes it hard to infer from results what the pattern of demand would be if the format generated higher or lower jackpots across the whole draw cycle rather than provide a sequence of lower followed by higher jackpots.

There are therefore limitations, in fact found in all time series modelling of lotto demand, to the style of study we have presented. On the other hand, the exercise has clear practical utility for operators. Moreover, results such as the tendency of regions with higher levels of schooling to purchase more tickets are relevant to issues of fairness when how to spend (and in particular where to spend) the proceeds of a lottery are debated.

Perhaps the highest priority in this area of lottery research would be to create data from pooling a series of cross-sectional surveys of households, or, better, from a longitudinal study of households, where behaviour was tracked draw-by-draw. Such a data set would be expensive to generate. Meanwhile, regional panel studies for other lottery jurisdictions are more feasible and it would be of interest whether results from Spain were replicated elsewhere.

References

Baltagi BH (2008) *Econometric analysis of panel data*. Wiley, Chichester

Combs KL, Kim J, Spry JA (2008) The relative regressivity of seven lottery games. *Applied Economics* 40: 35–39

Coughlin C, Garrett T (2009) Income and lottery sales: Transfers trump income from work and wealth. *Public Finance Review* 37: 447–469

Forrest D, Gulley O (2009) Participation and level of play in the UK National Lottery and correlation with spending on other modes of gambling. *International Gambling Studies* 9: 165–178

Forrest D, Simmons R, Chesters N (2002) Buying a dream: Alternative models of demand for lotto. *Economic Inquiry* 40: 485–496

García J, Rodríguez P (2007) The demand for football pools in Spain: The role of prices, prizes and the composition of the coupon. *Journal of Sports Economics* 10: 1–20

Garrett T, Coughlin C (2009) Inter-temporal differences in the income elasticity of demand for lottery tickets. *National Tax Journal* 62: 77–99

Ghent LS, Grant, AP (2010) The demand for lottery products and their distributional consequences. *National Tax Journal* 63: 253–268

Giacopassi D, Nichols MW, Stitt BG (2006) Voting for a lottery. *Public Finance Review* 34: 80–100

Goldstein H (2010) *Multilevel statistical models*. Wiley, New York

Greene WH (2011) *Econometric analysis: International edition*. Prentice Hall, London

Grote K, Matheson V (2011) The economics of lotteries: A survey of the literature. Research Paper 11-09, Department of Economics, College of the Holy Cross, Worcester, MA

Harrell FE Jr (2001) *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Springer, New York

Kitchen H, Powells S (1991) Lottery expenditures in Canada: A regional analysis of determinants and incidence. *Applied Economics* 23: 1845–1852

Mundlak Y (1987) On the pooling of time series and cross section data. *Econometrica* 46: 69–85

Oster E (2004) Are all lotteries regressive? Evidence from Powerball. *National Tax Journal* 57: 179–187

Perez L, Forrest D (2011) Own- and cross-price elasticities for games within a state lottery portfolio. *Contemporary Economic Policy* 29: 536–549

Price DI, Novak ES (1999) The tax incidence of three Texas lottery games: Regressivity, race and education. *National Tax Journal* 52: 741–751.

Wolfsen S, Briggs P (2002) Locked into gambling: anticipatory regret as a motivator for playing the national lottery. *Journal of Gambling Studies* 18: 1–17

doi:10.1111/pirs.12118

Resumen. Se ha elaborado un modelo de las ventas en 50 provincias españolas para 522 sorteos de un juego de lotería con un premio elevado. Un modelo de efectos aleatorios cruzados permite incluir perturbaciones aleatorias específicas para determinados sorteos y provincias en particular. El artículo explora cómo influyen los factores demográficos y socioeconómicos en el volumen de ventas, y presta especial atención a la relación entre las ventas y los ingresos *per capita* reales con botes de diferentes tamaños. Cuando los botes son bajos, las ventas aumentan a la par que el nivel de los ingresos provinciales. Pero esta relación desaparece o se invierte cuando los botes sucesivos elevan el bote, debido a que las ventas en las provincias más pobres son marcadamente más sensibles al tamaño del bote.

要約: スペインの50の県における522回の高額賞金のロトゲームの抽選からの売り上げをモデル化する。交差ランダム効果モデルは特定の抽選と特定の県に固有のランダムショックを考慮に入れる。本論文では、様々な賞金額での売り上げと一人当たり実質所得間の関連性を中心として、人口および社会経済要因がどのように販売金額に影響するかを分析する。賞金が低い場合、売り上げは県民の所得水準に伴い上昇する。しかし、賞金持ち越しにより賞金額が増加する場合、この関連性は失われるか逆転する。これは、貧しい県の売上高が賞金額に大きく反応するからである。