2003

Tax Incidence of the Pennsylvania Lottery: The Influence of Retail Outlet Location

Thomas Andrews
West Chester University of Pennsylvania, tandrews@wcupa.edu

Cynthia Benzing
West Chester University of Pennsylvania

TAX INCIDENCE OF THE PENNSYLVANIA LOTTERY: THE INFLUENCE OF RETAIL OUTLET LOCATION

Thomas Andrews and *Cynthia Benzing*, West Chester University

As of 2002, 38 states had adopted state lotteries to generate revenue to finance public programs. The Pennsylvania lottery raises funds for programs benefiting the state's older citizens. Lottery sales for 1999-2000 generated $670 million (39.3 percent of receipts) for programs such as the Property Tax/Rent Rebate program, the Mass Transit Program for the Elderly, and the Pharmaceutical Assistance Contract for the Elderly (PACE). Although Pennsylvania's lottery has been ranked by *International Gaming & Wagering Business Magazine* as the most efficient in the country based on the percentage of lottery revenue channeled to programs (www.palottery.com), concern remains about how the lottery affects the incomes of Pennsylvania's poorer citizens. The major argument against using lotteries as an excise tax is that the tax falls more heavily on lower income households.

Previous studies of the nature and causes of lottery tax incidence, including thorough analysis of demographic factors such as age, race, education, and unemployment, address the demand side issues and neglect the supply side. The state's role in the process has been largely ignored. The geographic distribution of lottery outlets has the potential to influence the regressivity of the lottery. If a state policy biases the location of outlets toward low-income areas, the lottery will tend to be more regressive. This is the first study to explicitly address the state's role in lottery regressivity.

Policies regarding commissions, fees, and outlet approval guidelines have the potential to affect the distribution of outlets, the distribution of sales, and the regressivity of the tax. State policies regarding the distribution of outlets vary, but most offer a commission for tickets sold plus a bonus for prizes awarded. In Pennsylvania, all retailers receive the same 5 percent commission irrespective of their location.

Some states allow any qualified dealer to sell tickets, while others, such as New York and Texas, will refuse applications in areas where a market is already well served. There is some evidence that, across a number of states, outlets per capita is greater in lower income areas Karcher (1989). The question remains, however, as to whether or not this is a completely demand driven phenomenon.

The Pennsylvania lottery consists of instant and online lottery products. Instant or "scratch" lottery games indicate winnings immediately. Instant tickets can be sold from machines and do not involve store personnel in administering them. In Pennsylvania, the instant lottery generates approximately 29 percent of revenues. Online games involve computers, communication networks at the site of play, and store personnel to place the bet. In Pennsylvania, the online games are the Daily Number, Big 4, Keystone Jackpot, Super 6 Lotto, and Cash 5.

This study uses the Suits Index to assess regressivity of the Pennsylvania online and instant lotteries under current policy. In addition, a sales model is used to determine how sales, and therefore the measure of regressivity, would change with a change in the distribution of outlets.

THE DATA

Monthly lottery dollar sales for the instant lottery and online lottery by outlet and zip code for 1999 were obtained from the Pennsylvania Lottery Commission. Demographic data by zip code were obtained from the U.S. Census data base (2000 figures were projections based on the last census; actual 2000 census figures were not available at the time of this survey). There are 1,470 zip codes in Pennsylvania. After eliminating 16 university or commercial zip codes with no income and little or no residential population, the sample size was 1,454.

THE SUITS INDEX

Suits (1977) developed an index to measure the level of tax progressivity or regressivity. Similar to the Lorenz curve and its Gini coefficient, the Suits Index provides a measure of the incidence of a particular tax. The Suits Index has been used in lottery studies by Clotfelter (1979), Clotfelter and Cook (1987), Hansen, Miyazaki and Sprott (2000), Hansen (1995), Price and Novak (1999), Price (2000), and Vaillancourt and Grignon (1988). It compares the accumulated percentage of a state's total lottery sales to the accumulated percentage of total household income. If the tax is proportional, the Lorenz curve will coincide with the diagonal line. If the tax is progressive, the Lorenz curve will lie below the diagonal. If the tax is regressive, the Lorenz curve will lie above the diagonal.

The formula for the Suits Index (S) divides the area between the Lorenz curve and the diagonal line by the area of the triangle below the diagonal. The equation is:

(1) $S = (K - L)/K,$

where K is the area of the triangle below the diagonal and L measures the area below the Lorenz curve. The Suits Index will lie between –1 and 1. When the Lorenz curve lies below the diagonal, area L is less than K and the index is positive. When the Lorenz curve lies above the diagonal, area L is greater than K and the index is negative.

To obtain the Suits Index, the percentage of total annual lottery sales by zip code and the percentage of total household income were accumulated across zip code from the lowest household income to the highest. Several studies have reported evidence that lotteries are regressive. Table 1 reports the Suits indices from a number of studies, along with estimates for Pennsylvania. Since each study relies on different states, lottery types, units of observation, measures of income, year, etc., the measures are not directly comparable; all but one, however, show the tax to be regressive. One additional factor that has not been addressed that may contribute to the differences in regressivity is each state's policy regarding outlet distribution.

SALES MODELS

The following section describes a model designed to estimate lottery ticket sales in each zip code in Pennsylvania. Ticket sales are classified as either instant (scratch and win) or online. Online sales require players to select a number that is recorded electronically and then wait for a drawing. While online sales occur for a variety of different games (pick 3, pick 4, etc.), all of these sales have been aggregated. In each of the models, the dependent variable is the total annual sales (instant or online) reported by retailers in each zip code for 1999.

The purpose of the model is twofold: (1) to determine the specific effects of demographic characteristics on sales and (2) to determine how the regressivity of the tax will change if the distribution of outlets is altered.

The regression results are based on a nonlinear model of the form:

(2) $sales = A\Pi x_i^{\alpha_i}$

which is essentially a Cobb-Douglas functional form where A is a constant, the x_i are the independent variables, and the coefficients, α_i appear as

Table 1
Suits Index Estimates for State Lotteries

Author	*Year Published*	*State*	*Lottery Type*	*Income Measure*	*Unit of Observation*	*Data Year*	*Suits Index*
Hansen, Miyazaki and Sprott	2000	OR	all	percapita	county	1995	0.01
Hansen, Miyazaki and Sprott	2000	FL	all	percapita	county	1995	–0.13
Hansen, Miyazaki and Sprott	2000	IN	all	percapita	county	1996	–0.06
Hansen, Miyazaki and Sprott	2000	CA	all	percapita	county	1996	–0.06
Hansen, Miyazaki and Sprott	2000	MN	all	percapita	county	1996	–0.17
Hansen, Miyazaki and Sprott	2000	NE	all	percapita	county	1995	–0.09
Hansen, Anne	1995	CO	instant	percapita	county	1989	–0.10
Clotfelter and Cook	1987	CA	instant	HHI	individual	1986	–0.32
Price and Novak	1999	TX	instant	Med HHI	zipcode	1994	–0.36
Price, Donald	2000	TX	instant	Percapita	county	1994	–0.13
Price, Donald	2000	TX	online	Percapita	county	1994	–0.05
Clotfelter and Cook	1987	MD	online	HHI	individual	1994	–0.42
Price and Novak	1999	TX	online	Med HHI	zipcode	1994	–0.26*
This study		PA	online	Med HHI	zipcode	1999	–0.45
This study		PA	instant	Med HHI	zipcode	1999	–0.48

*Average of multiple Suits indices

exponents. The independent variables include socio-demographic variables and the dealer density as a measure of lottery ticket accessibility. When linearized, this is the familiar double log model where the coefficients represent elasticities with respect to a change in the corresponding independent variable. The log-linear model must be estimated without including observations that have zeros for any of the variables. In this case, that would require a sacrifice of approximately 20 percent of the observations and make the results conditional on positive variables, possibly creating a bias. Rather than discard the many zero observations, the model is estimated in strict nonlinear form using an iterative process available in SPSS.

The advantage of this nonlinear form is that the first derivatives for any variable are not only a function of the level of the variable of interest, but also of the values of all of the other independent variables. So, for example, the change in sales predicted from adding one additional retailer will depend not only on the number of dealers but also on all of the other sociodemographic variables: income, population, etc.

$$(3) \quad \frac{\partial S}{\partial x_i} = \alpha_i A x_i^{\alpha_i - 1} \Pi x_j^{\alpha_j} \; \forall \; i \neq j$$

If $\alpha < 1$, the first derivative is negative and there are diminishing returns to that variable, while if $\alpha > 1$ there are increasing returns. In this model, the coefficients can be interpreted as elasticities.

Of specific interest in this study is the effect of adding dealers to any region. Because lottery tickets become more accessible as the number of dealers per square mile increases, it is assumed that there will be a direct relationship between the density of dealers in a zip code and sales. The only other published model, Hansen (1995), included population density as a proxy for ticket accessibility, but this formulation suffers from the fact that densely populated areas do not necessarily have more available outlets.

Including dealers as an independent variable results in a simultaneity problem. Sales depend on dealer density, but dealer density is also a function of ticket sales. The simultaneity problem causes bias in the estimate of the coefficient for dealer density, making it impossible to predict accurately the impact of additional dealers on sales. A Hausman endogeneity test showed with 95 percent confidence that the number of online dealers may be treated as an endogenous variable. The results for the instant ticket sales were inconclusive, but erring on the side of caution, this variable was also treated as endogenous. The simultaneity problem is addressed by adopting a two-stage approach. In ordinary least squares regression, the two-stage estimator is still biased, but becomes efficient. As the sample size increases, the estimated two-stage parameter becomes closer to the true parameter.

To correct for the simultaneity bias, the number of dealers was estimated using a regression with all of the socio-demographic variables from the sales function plus the number of potential lottery retail establishments in each zip code. Drug stores, liquor stores, grocery stores, and gasoline stations were classified as potential lottery outlets. The number of potential outlets should influence the number of dealers but not (at least directly) ticket sales. The predicted value from this first stage regression was used in place of the actual number of lottery outlets in the second stage regression to predict ticket sales.

The first stage model to predict the number of dealers is:

$$(4) \quad D = A \Pi x_i^{\alpha_i} R^{\alpha_j}$$

where R is the number of potential lottery retailers in each zip code and x_i are the sociodemographic variables from the sales equation. The model is nonlinear and similar to the second-stage sales regression. The nonlinear formulation is again useful in that it will give no predicted dealers in areas where there are no potential dealers.

Finally, the model was run in two versions. In the first, total annual instant lottery ticket sales per zip code was regressed against demographic characteristics and the number of instant lottery tickets outlets per square mile (predicted from the first stage model). In an analogous fashion, the second model regresses total online sales against demographics and the number of online outlets per square mile (again, the number of outlets is predicted from a first stage model).

Table 2 lists the variables and expected signs for each regression, and Table 3 shows the results from the first and second stage nonlinear regressions.

The results show a positive and nearly proportional effect of population on lottery sales. A 1 percent increase in population leads to a nearly 1 percent increase in sales. Increasing income decreases lottery sales, reinforcing the regressive

Table 2

Variables in the model (Zip code unit of observation)

Parameter	*Annual Sales of Instant lottery tickets (dependent)* *Annual Sales of Online Lottery tickets (dependent)* *Independent Variables*	*Expected Sign*
CONST	Constant	
MHHDI	Median Disposable Household Income	–
%ATLSTHS	Percent Of Population with at Least a High School Diploma	–
%UNEMP	Percentage of Labor Force Unemployed	+
%WHITE	Percentage of Population that is White	
APOP200	Population in 2000	+
MEDAGE	Average Age of Population	
%FEMALE	Percentage of Females	
ADSQM	Average number of lottery ticket (instant or online respectively) dealers Per Square Mile	+
Retailers	Number of potential lottery retailers (Drug stores, liquor stores, grocery stores, and Gasoline stations)	+

Table 3

Model Results (t-statistics shown below coefficient)

Dependent variable	*First stage Instant Dealers*	*Second stage Instant Sales*	*First stage Online Dealers*	*Second stage Online Sales*
Constant	42.132	27150.83	0.580	53.372
	0.867	0.494	0.767	0.494
POP2000	0.231*	**1.063***	0.215*	**0.936***
	10.861	59.144	9.186	43.622
MHHDI	–0.651*	**–1.207***	–0.349*	–0.151
	–12.624	–14.652	–6.166	–1.769
%ATLSTHS	–0.112	–0.141	0.047	**0.704***
	–0.991	–0.76	0.383	4.338
%UNEMP	0.0110	–0.001	0.183*	**0.519***
	0.394	–0.219	5.697	8.663
%FEMALE	–0.355	–0.868	–0.115	**–1.218***
	–1.252	–1.780	–0.353	–2.501
MEDAGE	0.410*	**1.561***	0.423*	**1.006***
	4.252	7.732	3.202	5.203
%White	0.164*	**0.906***	0.053	**0.135***
	5.152	12.932	1.628	3.440
DSQM		**0.097***		**0.263***
		9.310		24.349
Retailers	0.792*		0.815*	
	35.222		32.974	
Adj R-sq	.908	.815	.892	.825
n	1454	1454	1455	1455

*significant at at least 95 percent.

nature of the tax. Education and unemployment rates have a positive and significant effect on online sales, but are negative and insignificant in the instant sales regression. Age and the percentage of the population that is white increased lottery ticket sales, while the percentage of the population that is female decreased sales. The coefficient for dealer density (<1) supports the idea that an area could possibly become "saturated" with dealers, and the effect is more pronounced for instant dealers than for online dealers. In general, this study supports previous studies in that it indicates that lottery tickets sell better in areas where the population is older and less educated. It also finds that instant tickets sell better in poorer areas.

While the coefficients are useful for telling what will occur if one variable changes, the effects of policy changes are less obvious when many factors are involved simultaneously. Using this model, it is possible to predict the increase in sales that might occur in each zip code from adding an additional dealer. These predicted values can be analyzed to identify areas that are either over or under served. For example, poor areas with high dealer density may not generate the same sales that might occur in a wealthier area with a lower dealer density. The chart below shows the potential additional online sales that could be achieved by adding an additional dealer to each zip code. The additional sales are plotted against income. Points in the upper right represent high-income zip codes with high potential sales. The distribution of additional instant sales is similar, but in general added sales are lower. Only zip codes that have no unutilized potential retailers are plotted.

Chart 1 demonstrates that the marginal sales of an additional dealer are not equal across zip codes. In fact, the lowest additional sales also seem to correspond with areas with low income, and the highest sales occur in areas where income is above the state median of $38,000.

OPTIMAL OUTLET ALLOCATION

For a given number of outlets, the outlet distribution could be considered optimal if it maximizes lottery sales. In this case, that would mean applying the equal marginal principle familiar in economics. Sales will be maximized for a given number of outlets if the marginal contribution of an outlet to sales is equalized across districts. It obvious from Chart 1 that this is not the case in Pennsylvania. Projected sales from adding outlets vary from zero to almost $800,000.

If a state's policy regarding outlet allocation leads to a higher than optimal density in poorer areas, and vice versa, the tax will be more regressive than it could be, and at the same time the state will raise less revenue. Even a state that adopts "no policy" and allows demand and independent retailer decisions to be the sole determinant of supply will suffer at a minimum from lower than optimal revenue. Chart 2 illustrates the marginal sales of additional online outlets assuming that the state attempted to achieve the equal marginal principle. Though it is not realistic to expect the state to equalize the marginal outlet sales where there are insufficient potential retailers, the variance in marginal sales can be reduced substantially. Chart 2 shows the marginal sales that could be achieved by adding retailers in areas with higher than average marginal sales where potential outlets exist. The variance in potential sales is reduced from Chart 1, and revenue would increase by approximately $11 million or approximately 2.5 percent.

In this case, however, the consequence of failure to consider lottery outlet location is limited to a small amount of lost revenue. In fact the Suits Index increases slightly from -.45 to -.46 when the allocation of outlets is optimized. Chart 3 shows

Chart 1: **Additional On-line Sales by Adding One Outlet Where Possible**

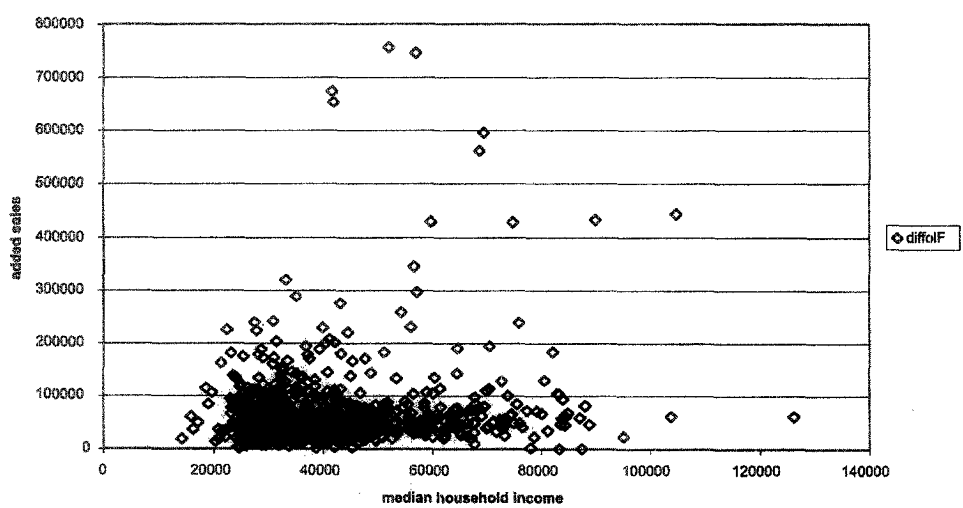


Chart 2: **Marginal On-line Sales by Income (As Close to Equal Marginal as Possible)**

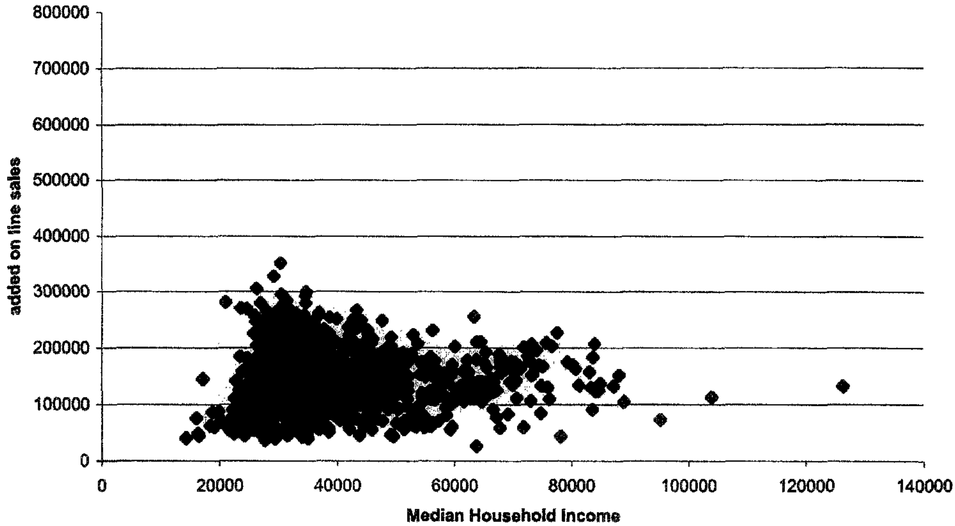
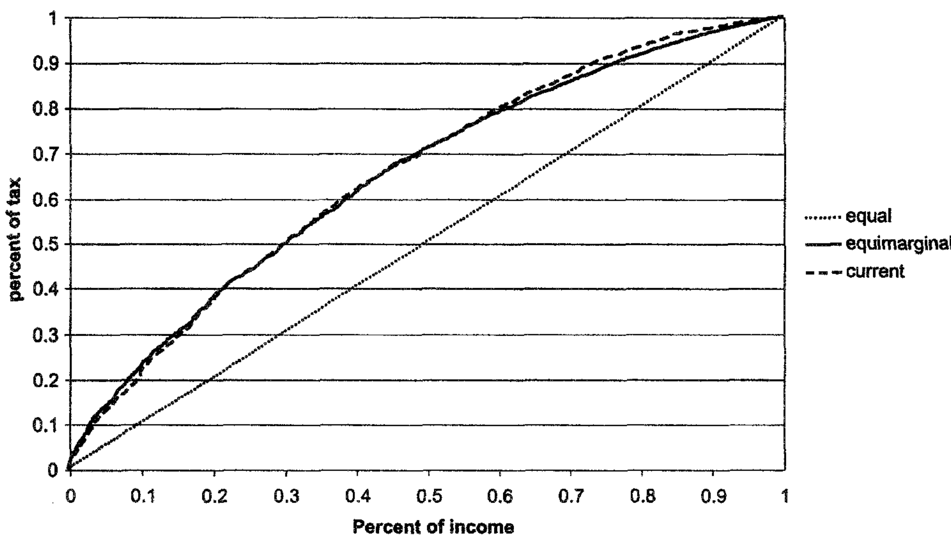


Chart 3: **Suits Index with and without Equal Marginal Principle**



the Suits curves for the current and the "optimal" outlet allocation. Since the curves cross, it is impossible to say which allocation is more regressive, but the chart below clearly demonstrates that the regressivity of the tax is not significantly altered by the redistribution of outlets.

CONCLUSION

State lotteries have become a significant source of revenue for state programs, and the evidence continues to mount that the burden of support for the programs financed by lottery sales falls disproportionately on poorer individuals. While this study reinforces the preexisting evidence that state lotteries are regressive, this study has also demonstrated that the role of supply in lottery tax incidence is surprisingly small. Furthermore, this study shows that while a reallocation of outlets might be able to increase revenue, it does not appear that there would be a significant redistribution of the tax burden. Further research may be necessary to determine whether these results also hold for other states. Since policies vary regarding approval of lottery outlet applications, it may be possible to correlate regressivity indices with policy parameters to determine significant factors. Another approach would be to develop a decision model for retail outlets where policy parameters factor in the decision.

Notes

1 The number of potential outlets was based on the number of establishments by SIC code obtained from the 1997 business patterns census. Some establishments were listed as existing in zip codes, which did not exist in the population census data base. These outlets were added to the count for the nearest zip code that existed in the population data base. The nearest zip code was determined using the MARKET function in ZIPFIP software. This software package will list the number of zip codes within a specified distance of an origin zip code. Using the unknown zip code as the origin, the specified distance was gradually reduced until only one other zip code remained.

2 Both models were estimated using an iterative procedure in SPSS. Starting values for the parameters were from the log-log version of the corresponding model.

3 All models including Hausman tests were estimated using the log-log and nonlinear functional forms. The nonlinear models performed better in all cases. The results from all models, however, are available from the authors on request.

References

Clotfelter, Charles T. On the Regressivity of State-Operated 'Numbers' Games. *National Tax Journal* 32 (December 1979): 543-548.

Clotfelter, Charles T., and Philip J. Cook. Implicit Taxation in Lottery Finance. *National Tax Journal* 40 (December 1987): 533-546.

Hansen, Ann. The Tax Incidence of Colorado State Lottery Instant Game. *Public Finance Quarterly* 23 (July 1995): 385-98.

Hansen, Ann, Anthony D. Miyazaki and David E. Sprott. The Tax Incidence of Lotteries: Evidence from Five States. *The Journal of Consumer Affairs* 34, 2 (2000): 183-203.

Karcher, A.J. *Lotteries*. Piscataway, New Jersey: Transaction Publishers, 1989.

Price, Donald I. The Income Redistribution Effects of Texas State Lottery Games. *Public Finance Review* 28, 1 (2000): 82-93.

Price, Donald I., and E. Shawn Novak. The Tax Incidence of Three Texas Lottery Games: Regressivity, Race, and Education. *National Tax Journal* 52, 4 (December 1999): 741-751.

Suits, Daniel B. Measurement of Tax Progressivity. *American Economic Review* 67 (September 1977): 747-752.

Vaillancourt, François, and Julie Grignon. Canadian Lotteries as Taxes: Revenues and Incidence. *Canadian Tax Journal* 36 (March-April 1988): 369-388.

APPENDIX

Hauseman tests t-stats shown below coefficients

Dependent variable	*First stage Online dealers/mi²*	*Second Stage Online Sales*	*First stage Instant dealers /mi²*	*Second Stage Instant Sales*
Constant	4.217E-20	1.291E+09	9.243E-19	1.314E+16
	3.549E-01	0.666	0.337	0.705
POP2000	–1.132	0.750	–1.101	0.751
	–30.367	46.640	–27.777	56.980
MHHDI	–0.115	0.061	–0.364	–0.793
	–0.684	.929	–2.234	–12.833
%ATLSTHS	6.618	–1.475	6.303	–2.948
	19.151	–9.332	17.602	–17.258
%UNEMP	0.778	0.209	0.624	–0.336
	5.659	4.563	4.709	–10.303
%FEMALE	13.577	–5.289	13.392	–7.486
	17.704	–13.845	16.193	–20.805
MEDAGE	–7.023	3.013	–6.564	3.900
	–24.032	17.584	–21.257	24.268
%White	–1.136	0.289	–1.096	1.204
	–11.780	9.000	–10.974	22.178
Retailers	1.885		1.806	
	28.803		26.070	
DSQM		0.237		0.453
		29.544		30.168
Predicted DSQM		**0.321**		**0.093**
		22.569		12.936
Adj R-sq	.661	.891	.611	.890
n	1454	1454	1454	1454

Model Summaries (nonlinear and log-linear) t-stats shown below coefficients

Dependent variable	**Instant Lottery Sales by Zipcode**				**Online Lottery Sales by Zipcode**			
Functional form	***Log-log***	***Log-log***	***Nonlinear***	***Nonlinear***	***Log-log***	***Log-log***	***Nonlinear***	***Nonlinear***
Endogeneity correction	***No***	***Yes***	***No***	***Yes***	***No***	***Yes***	***No***	***Yes***
Constant	13.082*	10.301*	13.018	10.209	3.684	1.475	4.863	3.977
	4.639	3.405	0.559	0.494	1.296	.456	0.593	0.494
POP2000	0.814*	0.804*	1.037*	1.063*	0.832*	0.831*	0.907*	0.936*
	40.925	37.267	61.277	59.144	41.116	35.879	47.316	43.622
MHHDI	–1.174*	–1.205*	–1.205*	–1.207*	–0.399*	–0.416*	–0.137	–0.151
	–8.576	–8.161	–15.625	–14.652	–2.869	–2.648	–1.800	–1.769
%ATLSTHS	0.619*	0.766*	–0.112	–0.141	0.721*	0.763	0.615*	0.704*
	1.790	2.043	–0.648	–0.76	2.037	1.890	4.352	4.338
%UNEMP	0.012	–.0129	–0.034	–0.001	0.136*	0.121	.426*	0.519*
	0.191	–0.187	–0.828	–0.219	2.085	1.625	7.891	8.663
%FEMALE	–1.221*	–0.272	–1.660*	–0.868	–0.335	0.363	–1.377*	–1.218*
	–1.938	–0.401	–3.898	–1.780	–0.531	0.500	–3.481	–2.501
MEDAGE	0.683*	0.638	1.634*	1.561*	1.276*	1.295*	1.174*	1.006*
	1.914	1.665	8.679	7.732	3.558	3.184	6.897	5.203
%White	1.057*	0.833*	0.960*	0.906*	0.0892	–.0460	0.080*	0.135*
	7.875	5.756	14.054	12.932	0.663	–.297	2.342	3.440
DSQM	0.293*	0.228*	0.131*	0.097*	0.408*	0.368*	0.289*	0.263*
	19.622	13.128	14.132	9.310	27.811	19.643	32.038	24.349
Adj R-sq	.724	.675	.831	.815	.804	.751	.857	.825
n	1106	1061	1454	1454	1050	1018	1454	1454