Journal Of Applied Economics and Policy

VOLUME TWENTY SIX, NUMBER ONE 2007

2006 KEA Best Student Paper

Refereed Papers

Journal of Applied Economics and Policy

VOLUME 26, NUMBER 1

2007

Published by the Kentucky Economic Association
At the College of Arts and Sciences
Eastern Kentucky University

The *Journal of Applied Economics and Policy* (formerly the *Kentucky Journal of Economics and Business*) is published by the Kentucky Economic Association at Eastern Kentucky University. All members of the Association receive the *Journal* electronically as a privilege of membership. All views expressed are those of the contributors and not those of either the Kentucky Economic Association or Eastern Kentucky University.

Future correspondence regarding articles submitted for publication in the *Journal* should be sent to Cathy Carey, Department of Economics, Western Kentucky University, 1906 College Heights Boulevard, Bowling Green, KY 42101. The submission fee is $20.00. A check should be made payable to the *Journal of Applied Economics and Policy*.

Communications related to membership, business matters, and change of address should be sent to Amy Watts, Long Term Policy Research Commission, 111 St. James Court, Frankfort, KY 40601.

The KEA's web page, www.kentuckyecon.org, is maintained by Tom Creahan, Morehead State Uninversity.

Table of Contents

The editor acknowledges the work of the following individuals who assisted with the review process for this issue of the *Journal*: Tom Creahan, David Eaton, Kenneth W. Hollman, Michael Jones, and Thomas G. Watkins.

Officers and Board of Directors of the
Kentucky Economic Association

President
Tom Creahan
Morehead State University

Secretary
Amy Watts
Long Term Policy Research Commission

Treasurer
Michael Jones
Governor's Office for Policy Research

President-Elect and Program Chair

Talina Matthews
Kentucky Office of Energy Policy

Board of Directors

2004 – 2007

David H. Eaton	Murray State University
Gail Hoyt	University of Kentucky
Talina Mathews	Kentucky Division of Waste Management
Bruce Sauer	LG&E Energy

2005 – 2008

Bruce Johnson	Centre College
Stephen Lile	Western Kentucky University
Chuck Martie	Governor's Office for Policy Research
Chris Phillips	Somerset Community College

2006 – 2009

Ali Ahmadi	Morehead State University
Jeff Florea	Madisonville Community College
Monica Greer	E.ON
Bob Houston	Eastern Kentucky University

KENTUCKY ECONOMIC ASSOCIATION
ROLL OF PRESIDENTS

James W. Martin	1975-76
Ray Ware	1976-77
Stephen Lile	1977-78
Dannie Harrison	1978-79
James McCabe	1979-80
Bernard Davis	1980-81
Frank Slensick	1981-82
Lawrence K. Lynch	1982-83
Clyde Bates	1983-84
Richard Sims	1984-85
Frank Spreng	1985-86
William Baldwin	1986-87
Richard Crowe	1987-88
Richard Thalheimer	1988-89
Lou Noyd	1989-90
Gilbert Mathis	1990-91
Claude Vaughn	1991-92
L. Randolph McGee	1992-93
Paul Coomes	1993-94
James R. Ramsey	1994-95
Ginny Wilson	1995-96
Bruce Sauer	1996-97
James Payne	1997-98
Mark Berger	1998-99
Martin Milkman	1999-00
Cathy Carey	2000-01
Manoj Shanker	2001-02
David Eaton	2002-03
Alan Bartley	2003-04
John T. Jones	2004-05
Brian Strow	2005-06
Tom Creahan	2006-07

Economic Analysis of Expanded Gambling in Kentucky

Darren Hobbs*

Abstract

Dwindling attendance at race tracks and the alleged drain on gambling tax dollars arising from out of state casinos have repeatedly brought forth the issue of expanding gambling possibilities in Kentucky. This study provides a pivotal first step in estimating the repercussions of instituting casino gaming by determining the demographic characteristics that best represent casino gamblers. Using data from the 1997-1999 National Gaming Impact Study and the 2003 Indiana Gambling Impact Survey, I conduct a binomial logistic model to estimate how educational attainment, income, and other gambling preferences are related to participation in casino gaming. In both data sets, college education and previous gambling behaviors are positively related to casino gaming while low household incomes were negatively related. With these coefficients, I compare the demographic information of Kentucky to that of Indiana and the nation to conclude that Kentuckians would be more likely than their counterparts to gamble at casinos if it were legalized.

I. Introduction

Declining profits at Kentucky racetracks and the ever present need for greater tax revenues have fueled several initiatives to institute casino-style gaming at Kentucky's horse tracks. Legislation, sponsored by the Kentucky Equine Education Project (KEEP), died in the Kentucky Legislature in March, 2006 when it was "banished" to dead-end committee debate (Alessi, March 23, 2006). The controversial issue of expanding gambling opportunities has garnered a number of fierce opponents. The National Coalition against Legalized Gambling (www.ncalg.org) correlates casino gaming with rising crime rates, bankruptcies, lost productivity, and social costs that far exceed the benefits of greater tax revenues. According to its literature, casinos tend to prey upon the poor and uneducated who are seemingly duped into pouring their entire paycheck into a slot machine. On the other side, KEEP (2005, 10) claims that their plan would provide hundreds of millions of additional dollars to fund education, healthcare, and economic development while simultaneously giving a boost to the horse industry which employs so many Kentuckians.

The implications of instituting casino gaming at Kentucky's racetracks include far too many facets for a single study to encompass. Before truly informed decisions can be made, one must be fully aware of the costs and benefits that accompany such a change. Aside from the question of how much additional tax revenue casinos can provide, legislators must evaluate the social costs associated with problem gambling and crime as well as the costs of providing transportation and security. As such, this study is in no way forecasting the social desirability of

*Undergraduate student, Transylvania University, Lexington, KY 40508. E-mail: dahobbs@transy.edu.

Kentucky casino gambling. Instead, the purpose is to take a scientific and objective approach to determining the characteristics and demographics that best represent casino gamblers. Using data from the 1997-1999 National Gaming Impact Study and the 2003 Indiana Gambling Impact Study, I conduct a scanning analysis to determine how educational attainment, income, and other gambling preferences are related to casino gambling. Through the use of a binomial logistic model, I estimate how these characteristics relate to casino gamblers. For instance, in both the national and Indiana data sets college education and previous gambling behaviors appear to be positively related to casino gaming. Having ascertained these values, the statistics can now be applied to Kentucky's particular set of demographics in an attempt to forecast if Kentuckians are more or less likely to embrace casino gaming than the average American. The results are not conclusive enough to fully define legislation, but do provide a pivotal first step in determining the validity of this endeavor, a motion that is sure to be a prime point of contention in Kentucky's next gubernatorial race (Alessi, August 21, 2006).

II. Literature Review

The initial debates over legalized lottery gaming began over four decades ago, and thus the available gambling research tends to focus more upon lotteries (Clotfelter and Cook, 1990, 106). Much of the lottery research revolves around a major tenet of economic analysis—the assumption that consumers are rational decision makers. The issue of legalized gambling questions this assumption when approached from almost any front. Many economists have pondered why rational consumers continue to "invest" in gambling when the expected payoffs are so low, especially in the case of lotteries. State lotteries, which have been legalized in 41 states and the District of Columbia, return between 40% and 60% of the initial investment while bingo, horseracing, and slot machines pay out 74%, 81%, and 89% respectively (Clotfelter and Cook, 1990, 107). All forms of legalized gambling, with the rare exception of extremely large lotto jackpots, have a negative expected return. If consumers are almost assuredly going to lose money over the long run, why do they continue to play?

The answer to this question encompasses a number of important aspects of gambling. Kearney (2005, 19) suggests that sheer entertainment value accounts for the discrepancy between the bet and the low expected payout. Morgan and Sefton (2000, 803) conclude that "when individuals account for the benefits from public good provision, funded from lottery proceeds, it becomes rational for risk-neutral individuals to participate in such a lottery." Thus it is not surprising to see the trend of "earmarking" gambling tax revenues for healthcare and education. Not only does it make the legislation more palatable for critics, but it also allows gamblers to feel that their losses are contributing to a greater cause. Even though Kearney (2002, 22) reports that a dollar of gambling profit specifically earmarked for education only increases expenditures by 60 to 80 cents, Morgan and Sefton's study shows that financing public goods does increase the willingness of potential gamblers to wager.

This issue of expected payouts and rational behavior highlights one important facet of gambling demographics, educational attainment. In regards to the lottery, Clotfelter and Cook (1990, 111) found that "lottery play is systematically related to social class, although not always as strongly as the conventional wisdom would suggest in this regard. The pattern is clear with respect to one indicator of social class: lottery play falls with formal education. The inclusion of

educational attainment as a factor questions whether all the consumers are maximizing utility rationally. In many ways, this revelation should not come as a surprise. As mentioned, lotteries provide a poor return on investment, a concept better grasped by the educated. Large multi-state lotteries like Powerball offer life-changing amounts of money at almost impossible odds. Baron and Kalsher (2005, 198) refer to this decision bias as the availability heuristic; this mental shortcut causes people to base decisions on how readily examples of an event can be brought to mind instead of considering the true probabilities. After viewing the publicity blitzes surrounding each winner, a rational person can more easily imagine winning a nine digit jackpot than conceptualize Powerball odds of 1 in 146,107,962.

It has been shown that lottery gaming falls with education, but what about casino gaming? Prior research tends to imply a finding in the opposite direction. Lotto and scratch offs, unlike certain types of casino gaming, require no knowledge of probability or strategy for enjoyment. Much to the chagrin of palm readers and fortune cookies, lotto drawings are completely random and no strategy can increase one's odds of winning. Casino games such as blackjack and video poker incorporate pure luck and a measure of skill (Eadington 1999, 178). In blackjack, a game where the typical house advantage is less than 1% (Eadington, 1999, 179), keen players can occasionally profit from the rare instances where the player actually has a slight advantage on the house. Knowing when such an opportunity arises requires a sharp memory, a strong knowledge of probability, and considerable experience with the game in question. Therefore, one can see how casino gaming is inherently more attractive to a more educated audience. This evidence supports the idea that educated individuals may be more likely to gamble at casinos and less likely to gamble lotto, but it in no way assesses the relative impact these forms of gambling have upon the groups.

However, many prevalent games within a casino accommodate the uneducated just as lottery games do. Unfortunately, these games also appeal to those prone to compulsive gambling. Research by Kearney (2005, 17) and the Kentucky Legislative Research Commission (2003, 21) have identified "instant games" as those most correlated to problem gamblers. Slot machines and roulette wheels, like scratch off tickets, are games of luck which also provide instant gratification and the opportunity for quick repeated play. Thus, it is plausible for casinos to draw the majority of their attendance from the educated but a majority of their revenues from compulsive gamblers.

Gambling by low-income households is often considered a highly regressive form of taxation. In her review of confidential Consumer Expenditure Surveys, Kearney (2002, 8) found that average lottery expenditures per quarter were uniform across low, middle, and high income households with respective expenditures of $125, $113, and $145. Consistent with these findings, Clotfelter and Cook (1990, 112) showed that the average lottery expenditure was nearly the same for households with income of $10,000 as for households with income of $60,000. Although this does demonstrate a highly regressive form of taxation, it tells us relatively little about the gambling behavior of the poor versus the rich. One would expect non-essential goods, like gambling, to be more prevalent in higher income households. On the other hand, low-income households may see gambling as a quick, although unlikely, solution to pressing financial situations. Society, being concerned with the welfare of the poor, often denounces regressive forms of taxation which place undue duress upon the low-income. However, the ideas

of casinos being more attractive to the rich and the perception of casino gaming as a regressive form of taxation are not mutually exclusive. The average expenditures of individuals who gamble may be similar, but this does not speak of the proportions of gamblers from each income level. It is entirely possible that another characteristic or set of traits jointly determine gambling behavior regardless of income.

Kearney (2002, 17) also shows that gambling in one capacity is significantly related to engaging in other forms of gambling, both legal and illegal. According to her study, "the introduction of a state lottery leads to a statistically significant 50.4 percentage point increase in the probability that an adult participates in gambling of any kind during the year." Although her study focused on the implications of legalizing state lotteries instead of casinos, her conclusions merit consideration. Kearney (2002, 20) found that the introduction of lottery gambling "might crowd in other gambling expenses, perhaps by reducing the "stigma" associated with gambling." Lottery, bingo, and horseracing are already legal forms of gambling in Kentucky, thus participation in these may influence the possible participation in casino gaming.

III. Methodology

The purpose of this study is to see how certain independent variables influence whether a person will attend a casino. With casino participation being a dummy variable, a dummy dependent variable model was selected for the regression. Although the linear probability model uses OLS and may be easier to interpret, the error term is inherently heteroskedastic and the results are not bounded between 0 and 1 (Studenmund, 2001, 36). Since the dependent variable is a dummy, the model should not predict the dependent variable to be above 1 or below 0. The binary logistic model avoids the problem of heteroskedasticity and ensures that the predicted outcome stays bounded between 0 and 1. This model utilizes maximum likelihood estimation and allows one to determine the likelihood that a person is in one category of the dummy dependent variable based on the values of independent variables. For example, a binary logistic model of car wrecks determines the increase in the probability of survival when wearing a seatbelt. This powerful model requires a sample of at least 500 data points, which is a requirement that both data sets in this study easily surpassed (Studenmund, 2001, 445).

After obtaining results for the national study, I studied the same model in the context of the Indiana survey. If the coefficients and levels of significance remain relatively consistent across the two surveys, the results are more robust and readily applicable. Showing similar results from the same model in two different data sets buttresses the validity of the model as well as the relative effects of the independent variables.

IV. Model

As previously stated, the intention of this research is to determine if Kentuckians are more or less likely to embrace casino gambling than the average American or Indiana citizen. The model, as shown below, is essentially one of demographics. Both the American Gaming Association (AGA) and the National Coalition against Legalized Gambling (NCALG) have produced statistics on the demographics of the gambling population, but these studies focus on average analyses. In particular, each side attempts to show that gambling does or does not draw

most of its consumer base from the ranks of the poor and uneducated. For instance, the AGA (2006, 33) produces literature which shows the average American casino player is a college educated, white-collar, 46 year-old man with an income of $56,663. This paints the picture that casinos draw mostly from middle and high income households, but the use of averages undermines the validity of these results. The situation could be as the numbers imply, or it is possible that a few men with unfathomable incomes could disguise the plight of thousands of players from below the poverty line. Thus, there is a need for an objective analysis to determine which demographics are characteristic of casino players. The model below is put forth to estimate the relationship between gambling at a casino in the past 12 months and demographic variables like educational attainment, income, and other gambling experience.

$$\ln\left(\frac{CASINO}{1-CASINO}\right) = \beta_0 + \beta_1 COLLEGE + \beta_2 POOR + \beta_3 TRACK + \beta_4 LOTTO + \beta_5 BINGO + \varepsilon_1$$

In this model, CASINO is a dummy variable for whether or not someone gambled at a casino in the past year. The dependent variable then measures the natural logarithm of the odds that a person gambled at casino in the past year and is useful as the first step in estimating how Kentuckians will respond to expanded gambling. The national data set asked participants whether they had ever gambled at a casino, gambled at a casino in the past year, and how often they gambled at a casino in the past year. Information on lifetime behavior includes many one-time gamblers and fails to provide the number of annual visits per year. Therefore, information on past year casino participation is best available measure of the dependent variable. Past year casino gambling also proved to be the most plausible on practical grounds as well. The Indiana survey only collected information on past year casino gambling, so using this variable made it possible to perform the scanning analysis. Determining whether a smaller or greater percentage of Kentuckians will actually engage in casino gaming opens the door for further research into the predicted prevalence of problem gaming and the corresponding social costs.

One of the most commonly questioned factors, and a factor heavily focused upon in this study, is that of education. COLLEGE, which is a dummy variable to identify people with some college education, is particularly important on two distinct fronts. The first is the research discussed earlier which has shown a link between educational attainment and lottery gaming. Prior research showed a negative correlation between education and lottery participation, but the same logic can be used to suggest a positive association between education and casino gaming, especially the games which reward those with greater knowledge. Secondly, education is important to the debate on whether casino gaming is a socially optimal policy since any policy that taxes the less educated disproportionately is sure to garner heavy criticism.

The second independent variable, POOR, is a dummy variable to identify poor households. POOR tests the conventional wisdom that gambling is most attractive and most problematic for low-income households. Gambling, like other non-essential normal goods, should be consumed in lower proportions by the low-income. However, there is reason to doubt this and reason to expect different patterns of participation between casino and lottery gaming. Although the expected payoff of lottery games is significantly lower than any other form of gambling, the "cost" of playing the lottery is relatively low. The cost of gambling at a casino or horse track is raised considerably by the price of admission, parking, programs, lodging, and especially transportation. On the other hand, lottery games are available at almost every gas

station or supermarket and can be purchased in the course of routine errands. In this respect, lottery gambling should be more attractive than casino gaming to the poor.

The final three independent variables account for other gambling experience in the past 12 months. Gambling in other capacities should serve as a strong indicator that one would engage in casino gambling were it to become available. LOTTO, TRACK, and BINGO are dummy variables equal to one if the person gambled in a lottery, at a horse track, or at bingo, respectively, in the past year. Just as lottery and casino draw from different demographics, so too do handicappers and bingo players. By measuring the independent effect of each form of gambling on the odds of casino gambling, we can assess which of the three types is most critical in estimating casino participation. Of the three types of gambling tested (lottery, horses, and bingo), I expect gambling on horseracing to have the strongest association with casino gaming. The reason for this follows the same logic as the argument for the positive relationship between casino gaming and educational attainment. In much the same way as a skilled blackjack player can determine when the odds are in his favor, an experienced and knowledgeable handicapper can greatly tilt the odds in his favor. Anyone who consistently gambles on horses has some knowledge of probability and strategy, thus it is not a stretch to assume that they would be more likely than others to feel comfortable playing casino games.

As previously mentioned, lottery and bingo require none of the knowledge or skill needed for some casino games, but I still expect a positive relationship between these variables and casino gaming. Gambling on bingo and lottery, two forms of gambling with the lowest expected returns, could demonstrate a risk-loving trait which would be conducive to casino gaming. However, some individuals might play the lottery just to support education and others may play bingo for the sole benefit of a church youth group, but the desire for personal gain or entertainment is likely the prime motivation for most of these gamblers.

V. Data

Although it would seem most prudent to estimate the regressions on a data set from the state of interest, the data available for Kentucky could not be reasonably applied to the model. The most recent study, the Kentucky Legislative Research Commission's Report 316 on Compulsive Gambling, will be used to forecast the implications of expanded gambling on Kentucky, but the included information was not applicable to the regression analysis. In order to correctly forecast the potential behavior of Kentuckians, the data must be drawn from respondents whose states have legalized casino gaming.

The first data set used, and the one most commonly referenced in previous research, is the National Gambling Impact Study. The national survey of demographics, gambling behavior, and opinions, was conducted between 1997 and 1999 by the National Opinion Research Center at the University of Chicago. This 2,947 person survey is a conglomeration of results from a digit dial sample and smaller samples of gambling and non-gambling patrons. The questionnaire, which included hundreds of questions, is quite in-depth, but the responses were generally recorded categorically rather than numerically. This categorical data set is not conducive to a standard OLS regression analysis, but can be used to estimate the binary logistic model employed in this paper. Building upon the conclusions of prior research, I had a strong

inclination as to the signs of the variables but felt that a scanning analysis would be most appropriate for this particular study. Comparing the results from this broad data set to a narrower one serves as a strong test of the validity of the results.

The second data set is the Indiana Gambling Impact Study (IGIS) which was conducted in 2003 as a joint venture of Indiana University Purdue University at Indianapolis (IUPUI) and the Indiana Criminal Justice Institute. This 801-person, digit dial sample has a significantly smaller scope than the National study; however, being a neighboring state, the demographics and opinions of the Indiana population are perhaps more likely than the national sample to reflect those of Kentucky citizens. Also, the Indiana survey is six years more recent than the national study; in the past decade, there has been a strong proliferation of casinos, racinos, and riverboats throughout the country. Thus, the attitudes and behaviors reflected by the IGIS display a more current perspective.

Table 1 shows the frequency distributions for variables measuring gambling behavior, educational background, and income in the national and Indiana surveys. Overall, lottery gaming is by far the most popular with casinos being a distant second. The distributions for educational attainment and income show that the samples tend to be characterized by slightly wealthier and more educated citizens than Census averages in Table 4 would suggest, but the relative proximity of these numbers to their true values demonstrates that conclusions from this data can be applied to other populations.

Table 1. National and Indiana Summary Statistics

		National	Indiana
Gambled Casino in Past Year?	Yes	34.1	17.8
	No	65.9	82.2
Gambled at Horse Track in Past Year?	Yes	10.8	4.7
	No	89.2	95.3
Gambled on Bingo in Past Year?	Yes	6	5.5
	No	94	94.5
Gambled on Lottery in Past Year?	Yes	56.3	39.8
	No	43.7	60.2
Education beyond High School?	Yes	59.1	63.1
	No	40.9	36.9
Income below $24,000*?	Yes	33.6	20.8
	No	66.4	79.2

*Income below $30,000 for Indiana.

Finally, in a regression where the dependent and all independent variables are dummy variables, the chance for problems arising from multicollinearity can be high. Intuitively, I was particularly apprehensive about the correlation between each of the three types of gambling and the possible correlation between education and income. To dispel these questions, I estimated a correlation matrix which shows the correlations between each of the independent variables. Overall, COLLEGE and POOR were the most highly correlated variables, but the correlation value of .264 is not high enough to warrant real concern. Of the gambling variables, LOTTO and BINGO were the most correlated with a value of -.076. Even though the relatively low

correlations between each pair of independent variables do not rule out the presence of multicollinearity entirely, but they disprove the existence of a first-order correlation.

The categorical, rather than numerical, recording of data required recoding for all variables. In both surveys, the variables for CASINIO, TRACK, LOTTO, and BINGO were carefully recoded to correct recording incongruities and increase study validity. The wording of the questions and the manner in which the studies were conducted made these changes necessary. It appears that anyone who answered "no" to "Have you ever gambled in a Casino?" was directed not to answer the question about participation in the past year. Although it is apparent that anyone who has never gambled in a casino could not have possibly gambled at one in the past year, the researchers coded the responses from those who had never gambled as missing responses in the past year question. Correcting for this discrepancy nearly doubled the sample size and rightfully added those who had never gambled to the ranks of those who had not gambled in the past year.

In the Indiana Survey, the CASINO variable required additional recoding, but the action had a strong theoretical rationale. In 2003 there was and still is a strong dichotomy between the gambling opportunities in the Northern and the Southern portions of Indiana. Southern Indiana casino gamblers are more likely to flock to the riverboats along the Ohio River, which are conveniently placed to induce a larger share of Kentucky gamblers. In the North, most of the gambling opportunities come from Illinois casinos placed on the northwestern border between the two states. The survey asked questions about these two types of gambling separately, but for the purposes of this research, gambling on a riverboat and in Illinois casino represent the same behavior. As such, in the Indiana regressions, a person who had gambled in either of these two capacities was assigned a value of one, while all others were coded as zeros.

In this model, the dummy variable, COLLEGE, was coded as one for anyone who had pursued education beyond a high school degree and zero for all others. Theoretically, when people pursue higher education, they are more likely to learn about probability, which helps them understand the nature of some casino gaming. Also, the categorical recording of the educational attainment data was slightly different between the two studies; as such, this distinction allowed for the most consistency across the two samples used in the scanning analysis.

Finally, the distinguishing annual income for the POOR variable was $24,000 - $30,000. In what was likely a way of increasing response, the income variable in both surveys was recorded in restrictive categories. Ideally, income should be recorded numerically so that the differences in participation could be observed across the whole range of incomes. With this data unavailable, the lowest income group for each survey was chosen in order to represent the low-income respondents. In the national survey, this income group included all those with incomes between $0 and $24,000 while the Indiana category included incomes between $0 and $30,000. Both categories are reasonably close to the poverty line income of $20,000 for a four-person household. POOR has a value of one in the national survey if the household reported income between $0 and $24,000 and a value of one in the Indiana survey if the household reported income between $0 and $30,000.

VI. Results

The results from the national test turned out as anticipated. Each of the independent variables was significant at a .05 level of significance, and the coefficients demonstrate the projected signs. Table 2 shows the coefficients, significance levels, and the exponential of each coefficient. This final column, Exp(β), is the most significant for analysis and forecasting. In a binary logistic regression, the left hand side of the equation uses the natural log of the dependent variable. Thus, the exponential of the coefficient produces a value that shows how each independent variable increases or decreases the odds of a positive value for the dependent variable. A value greater than one denotes a positive relationship while a value less than one displays a negative relationship.

Table 2. National Survey Results.

	Coefficient	Significance	Exp(β)
COLLEGE	0.350	0.002	1.419
LOTTO	1.186	0.000	3.275
TRACK	1.289	0.000	3.629
BINGO	0.957	0.000	2.603
POOR	-0.246	0.036	0.782

The exponential of the estimated coefficient measures the effect of the independent variable on the likelihood of casino gambling. The exponential of the coefficient on COLLEGE suggests that individuals with education beyond high school are 41.9% more likely to gamble at a casino than those with lesser educational attainment. Households with incomes below $24,000 are shown to be 22% less likely than higher-income households to gamble at a casino[1]. The results tend to reject the conventional wisdom that casino gaming preys upon the poor and uneducated, but these results only show the relationship between the demographics and gambling at a casino at least one time in a year. Poor and uneducated individuals may be less likely than others to gamble at a casino in general, but it is entirely plausible for these individuals to be the most susceptible to problem or compulsive gambling.

As had been forecast, gambling at a horse track in the past year was the most significant determinant of casino gaming with lotto and bingo participation falling second and third respectively. Track gamblers are 262% more likely than non-track gamblers to gamble at a casino while lottery and bingo players are 227% and 160% more apt to play casino games than their non-playing counterparts. This reinforces the intuition that prior knowledge of gambling strategy translates into participation in other games that rely upon this same knowledge. However, as the figures from Lotto and Bingo suggest, any previous gambling experience contributes to a greater likelihood of gambling at casino games.

[1] In a separate regression, the variables of COLLEGE and POOR were tested alone to see if they remained consistent and significant without the driving force of the other strong variables. The analysis returned results similar to the primary regression, thus further proving the significance of COLLEGE and POOR on their own accord.

In a binary logistic regression, the Omnibus Test of Model Coefficients tests the overall validity of the model by determining whether the independent variables, when considered as a set, can significantly explain variations of the dependent variable more accurately than the constant term alone. This relationship had a Chi-square value of 281.633 with a significance of .000. Thus, at the .001 level, the independent variables as a group are statistically significant. A final test of the overall model, the Hosmer and Lemeshow Test, is a logistic test of goodness of fit. In stark contrast to most tests, a significance level of .05 or greater demonstrates that the theoretical model describes the variance of the dependent variable in a statistically significant way. The significance value for this regression was .774, thus the model demonstrates a more than adequate goodness of fit.

According to theory, TRACK, LOTTO, and BINGO are related to casino gaming because one, people who enjoy gambling in these capacities would likely embrace casinos, and two, gambling knowledge learned through other forms could be applied to casino games. Both theoretical underpinnings are important, but the analysis as it stands cannot distinguish between the two effects. In order to determine which is more important, I combined the columns for TRACK, LOTTO, and BINGO into a single column of GAMBLER that assigned a 1 to anyone who had gambled in any capacity in the past year. In essence, this analysis was conducted to determine if the act of gambling itself is more important than the differences between the types of gambling. When estimated, the value for GAMBLER was as strong and significant as the three variables separately. The statistical significance of GAMBLER shows that the act of gambling is an important determinant, but the knowledge of probability and strategy that accompany specific forms of gaming is more important in predicting the acceptance of casino gambling.

The results show a positive relationship between casino participation and higher education, but the theory upon which this is based implies that forms of gambling lie on a continuum ranging between pure luck and a combination of luck and skill. According to theory, individuals with greater education will gravitate towards the casino games where they can apply their knowledge to improve the odds of winning. Also, people with more education are more likely to realize that the expected return of casino gaming exceeds that of the lottery. In another regression, LOTTO was set as the dependent variable and regressed against the independent variable of COLLEGE. The analysis concluded that a statistically significant negative relationship exists between the variables. Individuals with education beyond high school were 17% less likely to have played the lottery in the past 12 months. This result further reinforces the theory relating higher education to casino gambling.

Having found the expected results using a national survey, I then applied a similar model to the Indiana survey in order to test the validity of the results. The values and significance levels are summarized in Table 3.

Table 3. Indiana Survey Results

	Coefficient	Significance	Exp(β)
POOR	-0.782	0.019	0.457
COLLEGE	0.494	0.042	1.639
BINGO	1.364	0.000	3.911
TRACK	1.716	0.000	5.560
LOTTO	1.608	0.000	4.991

Just as before, the variable for past year horse gambling, TRACK, was the one most strongly associated with casino gaming. LOTTO and BINGO were second and third, respectively, just as they had been ranked in the national survey. Again, COLLEGE was expectedly positive, but the Indiana results showed that education beyond high school increases the likelihood of gambling in a casino by 63.9%. Indiana households with incomes below $30,000 were 55% less likely to gamble casino compared to the 22% reduction in likelihood shown in the national sample. A portion of this change could be attributed to the change in the upper bound of the income range from $24,000 to $30,000. However, when adjusted for inflation[2], $24,000 in 1997 is approximately $27,500 in 2003, so the change resulting from this sample discrepancy could be negligible. The test of the overall model, the Omnibus Test of Model Coefficients, had a Chi-square value of 111.841 and significance of .000. At the .001 level of significance, this set of independent variables is significantly related to the dependent variable.

Finding comparable results from this narrower and more recent study suggests the soundness of the theoretical model. The effect of each independent variable appeared to be more pronounced in the Indiana survey, but the variables maintained their positions of strength relative to each other. The similarity of the coefficients, levels of significance, and the likelihood of casino gambling for the Indiana and national samples supports the legitimacy of the theoretical model and the relationships between the variables. The gambling characteristics which influence casino gambling remained consistent over the past decade and have been identified in two distinct surveys. The significance of finding parallel results in the Indiana survey can not be understated. The Indiana results not only support the notion that the theory models gamblers in general, but also suggest that the theory models gamblers with qualities characteristic of Kentuckians.

VII. Forecast for Kentucky

The data and regressions have shown that higher education and participation in other forms of gambling are significantly and positively related to casino gaming while having a low income makes one less likely to play. With this in mind, how does Kentucky compare to the Indiana and the nation as a whole? Table 4 contains the percentage of people from each area that have pursued higher education, gambled in varying forms over the past 12 months, and have incomes below $25,000. The figures for educational attainment and income were gathered from

[2] Adjusted for inflation using the Bureau of Labor Statistics Inflation Calculator. www.bls.gov

the 2000 Census, and the gambling participation rates are from the three separate surveys of gambling behavior[3].

Table 4. Gambling, Educational Attainment, and Income in Kentucky, Indiana, and the Nation.

	Kentucky	National	Indiana
Education	40.6	50.9	44.9
Past Year Casino	15.1	34	17.8
Past Year Bingo	22	6.0	5.5
Past Year Horses	16	10.9	4.8
Past Year Lotto	41.7	56.3	39.8
Income under $24,000	37.7	28.7	27.8

At 40.6%, Kentucky ranks several percentage points below both the Indiana and national averages for adults with some education beyond high school. The regression results indicate that obtaining higher education increases the likelihood of playing casino games by 41%-63%. Also note that according to the Census, Kentucky has a greater population of individuals with incomes falling below the $24,000. Having a low-income has been shown to decrease potential for casino gaming somewhere between 18% and 55%. In these two respects, it appears that Kentuckians as a whole would be less likely to gamble at casinos.

The participation rates for horse and bingo gambling paint a decisively different picture. Kentuckians embrace racetracks and bingo halls with more vigor than the average American and Indiana citizen. Participation in these activities is three to four times more prevalent in Kentucky than in Indiana. Gambling on bingo increases the likelihood of casino gaming by 160-291% while racetrack gambling increases the likelihood by 262-456%. Although the populations of racetrack and bingo gamblers are smaller than the populations of less educated and low-income individuals, the strength of these variables suggests that the impact of Kentucky's pre-existing gambling population would cause Kentuckians to be more likely than the average American to gamble at casinos.

In addition, saving the horse racing industry has been a primary objective of proposed casino gaming legislation. To many, horses are inseparably tied to the image and history of Kentucky as well as the livelihood of its citizens. According to KEEP, the 80,000 – 100,000 jobs and $4 billion generated by the Kentucky horse economy have been jeopardized by the encroaching presence of expanded gambling from out-of-state sources (Kentucky Equine Education Project, 2006, 2). Economists may scoff at arguments based on saving jobs, but these statistics tend to weigh heavily upon the voting public. Such a policy may not "save" horseracing and its illustrious history in Kentucky, nor may it be a socially beneficial program, but horse enthusiasts may perceive gambling losses at Kentucky casinos as a form of public good financing. This harkens back to the research by Morgan and Sefton (2000) which found that individuals are more likely to gamble and lose greater sums if they feel they are financing a

[3] The Kentucky participation rates come from a 1200 person digit dial survey conducted by the Kentucky Legislative Research Commission in 2003.

public good. Even though these benefits may never come to fruition, anyone who feels, correctly or incorrectly, that casino gambling saves horses, jobs, history, education, or healthcare would be more inclined to participate and lose greater sums. The impact of this "public good" mentality would certainly be more pronounced in the population of racetrack gamblers, but the implications would likely spill over all demographics. Even though the figures for education and income speak otherwise, Kentucky's large gambling population combined with the seemingly sacrosanct standing of horseracing suggest that Kentucky casinos would stay busy.

VIII. Conclusion

This research is an important first step in a long series of topics that deserve consideration prior to the implementation of any casino gaming law. Opponents of expanded gambling often cite the costs of divorces, crime, and lost productivity which accompany problem gambling and negatively impact all citizens. This study has concluded that Kentuckians would be more apt to gamble at casinos, and thus a greater incidence of problem gambling is a concern. A much higher incidence of compulsive gambling could skew the cost-benefit analysis against casino gambling.

A second source of concern is the effect that Kentucky casinos would have on surrounding businesses. Will the rising tide of casinos lift the boats of local hotels, restaurants, and businesses? Or will the presence of casinos "cannibalize" entertainment dollars in the region? Showing the demographics that influence casino gaming is the first step in this research, but the expenditures of these patrons outside the casino walls must now come into question.

This paper has contributed two important points to the policy debate surrounding Kentucky's potential implementation of casino gambling. The first is the analysis of the demographics that generally characterize casino gamblers. Higher education, along with the greater ability to understand probability and gambling strategy, tends to lead individuals towards gambling outlets with higher expected returns (casinos) rather than low-return games (lotteries). Casino gambling, being a luxury good, has also been shown to be most attractive to the higher income brackets. Finally, a person's previous gambling experiences, especially horseracing, appear to be the most predictive determinants of casino gambling. The additional knowledge of gambling or simply the risk-loving behavior displayed through these activities increases the likelihood that an individual will gamble in a casino venue as well. Secondly, I have concluded that Kentuckians are more likely than the average American to participate in casino gambling, if instituted. Having applied the aforementioned relationships to the specific characteristics of Kentucky, it becomes apparent that Kentucky's large gambling population and the notion of public good financing would fuel a strong casino participation rate.

References

Alessi, Ryan. "Gambling likely to be issue in '07," *Lexington Herald-Leader*, Aug. 21, 2006.

Alessi, Ryan. "Casino proposal appears dead for this session," *Lexington Herald-Leader,* March 23, 2006.

American Gaming Association, "State of the States 2006." www.americangaming.org/assets/files/2006_Survey_for_Web.pdf

Baron, Robert and Micheal Kalsher. *Psychology: From Science to Practice*. Pearson Education, Inc., 2005.

Boardman, Barry, Jack Jones, John Perry, and Murray Wood. "Compulsive Gambling in Kentucky." Kentucky Legislative Research Commission Research Report No. 316, 2003.

Clotfelter, Charles T. and Philip J. Cook. "On the Economics of State Lotteries," *Journal of Economic Perspectives,* 4(4), Fall 1990, 105-119.

Eadington, William R. "The Economics of Casino Gambling," *Journal of Economic Perspectives*, 13(3), Summer 1999, 173-192.

Indiana Criminal Justice Institute. Gambling Study, Unpublished Research Study, 2003.

Kearney, Melissa S. "State Lotteries and Consumer Behavior," NBER Working Paper No. 9330, 2002.

Kearney, Melissa S. "The Economic Winners and Losers of Legalized Gambling," NBER Working Paper No. 11234, 2005.

Kentucky Equine Education Project. "The Facts on Gaming." 2005.

Kentucky Equine Education Project. "The Purpose, The People, & The Programs." 2006.

Morgan, John and Martin Sefton. "Funding Public Goods with Lotteries: Experimental Evidence," *The Review of Economic Studies*, 67(4), October, 2000, 785-810.

National Coalition against Legalized Gambling. www.ncalg.org.

National Gambling Impact Study Commission. Washington, D. C. *GAMBLING IMPACT AND BEHAVIOR STUDY, 1997-1999*: [United States] [Computer File] ICPSR02778-v1. Chicago, IL: National Opinion Research Center, [producer], 1999. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, [distributor], 2002.

Studenmund, A. H. *Using Econometrics: A Practical Guide, Fourth Edition*. Addison Wesley Longman, Inc. 2001, 442-448.

Organic and All Natural: Do Consumers Know the Difference?

Jeffrey Anstine*

Abstract

In 2000 the U.S. Department of Agriculture established new requirements for products labeled organic. The new rules were due to in part to consumers' confusion and misinterpretation of the word organic. This paper examines consumers' willingness to pay for dairy products, milk and yogurt, labeled natural and organic, compared to their counterparts that are not. As expected, households are willing to pay a significant premium for organic milk. Consumers are also willing to pay more for yogurt labeled 'all natural' and yogurt labeled 'organic' compared to yogurt without these labels. However, there is no statistically significant difference between consumers' willingness to pay for yogurt that is all natural and yogurt that is organic. We would expect consumers to be willing to pay more for organic yogurt than all natural yogurt since all natural yogurt may contain bovine growth hormones and organic yogurt cannot. If consumers do not know the difference in the terms organic and all natural, they may be willing to pay the same premium for all natural and organic yogurt compared to yogurt without these labels.

I. Introduction

Over the past few decades biotechnology and other advances in farming have enabled agriculture to vastly increase the amount of food produced. While gene manipulation has improved crop yields and the use of hormones in livestock has increased animal output, these events have raised consumer questions about the safety of some basic food products. As a result, more people are buying food that is perceived to be healthier than other similar products. When possible, consumers are opting to buy food that is labeled 'all natural' or 'organic.'

The organic industry, including organic dairy products, grew approximately 20 percent a year throughout the 1990s (Organic Trade Association, 2005). The market for organic dairy products has grown dramatically over the last decade due primarily to concerns over the use of hormones to increase milk output in cows. "Horizon Organic, the biggest organic dairy company in the country, added 64 organic dairy farmers in 2006 for a total of about 350, and about 230 more are in transition, said Sara Unrue, a spokeswoman." (Martin, 2007). Horizon's sales have grown an average 127 percent per year since 1993.

While some consumers may view all natural and organic as similar, they are different. While organic foods are all natural; all natural foods are not necessarily organic. The requirements for organic foods are more stringent than the requirements for all natural foods.[1] Organic milk and yogurt have to be from cows that are fed organic grain and not treated with

[1]In "Organic Food Standards and Labels: The Facts," the United States Department of Agriculture describes the standards that must be met for food to be considered organic. The USDA states that natural and organic are not interchangeable.

hormones and have to meet other stringent requirements. All natural products, such as yogurt, just cannot contain synthesized ingredients.[2]

If the goal is to eat healthier food, then consumers should be willing to pay more for products that are higher in their level of 'naturalness.' This paper uses the hedonic price technique to determine if consumers are willing to pay more for all natural products compared to those that are not and more for organic foods than for those foods that are just all natural.

II. Literature Review and Hedonic Model

Most of the research examining household buying decisions for organic products has been on produce, not packaged goods. Park and Lohr (1996) found that consumers were willing to pay a premium of 25 - 30% for organic broccoli, romaine lettuce and carrots. They also predicted that the growth in organic markets would initially be caused by consumer demand, which would then increase supply. Thompson (1998) summarized earlier work regarding purchases of organic fruits and vegetables, which mostly relied on self-reported data from consumers. Results were generally conflicting regarding which variables are important in determining who purchases the more expensive organic produce. While some studies found that households with a higher income, people with more education, females, and married couples generally were more likely to pay a premium for organic fruits and vegetables, these results did not hold for all studies.

Using a telephone survey, West et al. (2002) found that Canadian consumers perceived genetically modified negatively compared to their non-modified counterparts. In addition, in response to hypothetical questions consumers were willing to pay a premium for foods that were perceived to be healthier. After controlling for perceived healthiness, they were not willing to pay more for organic foods. Most recently, Dhar and Foltz (2005) examine the benefits to consumers in Wisconsin from rBST-free and organic milk compared to non-organic milk containing rBST. Milk containing rBST is produced using bovine growth hormones that increase milk production in cows. Thus, milk labeled rBST-free does not use these hormones. Organic milk, in addition to not containing rBST, is also produced from cows that have not been fed grain that has been grown using pesticides or herbicides. Non-labeled milk contains rBST and is not organic. Thus, there are three levels of naturalness. Organic is the most natural, rBST-free is the next most natural and non-labeled is the least. The authors find that consumers benefit from being able to buy milk based on these labels.

Ippolito and Mathios (1993) examine the usefulness of the Food and Drug Administration's Nutrition Labeling and Education Act of 1990. The Act was intended to force firms to provide better information on labels about the health and nutrition of their products. While consumers can be confused about the nutrition of foods, the authors find that the mandatory labeling could also reduce helpful, non-deceptive claims by firms and thus may adversely impact consumer knowledge of foods.

[2] Also see the Organic Trade Association for more information on the terms natural and organic.

Other research has used the hedonic approach to determine how consumers value certain product characteristics. The majority of the literature has focused on housing and durable goods markets. In a few cases the hedonic technique has been used to see what attributes of non-durable goods consumers value. Stanley and Tschirhart (1991) found that consumers in Portland Oregon are willing to pay more for cereals with higher vitamin content and additions like dried fruit. Nimon and Beghin (1999) found that consumers are willing to pay a premium for apparel made with organic fibers. Anstine (2000) showed that plastic garbage bags with handles sell for a premium, but that bags made with recycled material do not. This paper follows the approach of these authors and uses the hedonic price technique to determine the premium, if any, that consumers are willing to pay for all natural and organic yogurt and milk.

The hedonic model can be applied to a market for any differentiated product, here dairy products.[3] Consumer utility, U, is a function of a composite good, X, Milk, M and yogurt, Y, and taste parameters, T, such that $U_i=u_i(X_i, M_i, Y_i; T_i)$. Thus, each consumer has a family of indifference curves representing their tradeoff between the attributes of milk and yogurt. An individual maximizes utility subject to a budget constraint, $\sum P_i*X_i+ P_M*M_i +P_Y*Y_i=I_i$, where P_i is the price of the composite good, P_M is the price of milk, P_Y is the price of yogurt and I is income. Constrained optimization yields a set of demand functions where $Y_i=y_i(P_Y, P_M, T_i, I_i)$ and $M_i=m_i(P_Y, P_M, T_i, I_i)$.

Firms offer milk or yogurt with different characteristics in order to satisfy the various tastes of consumers. Each combination of attributes carries a different price that reflects the marginal cost of producing each attribute and consumers' willingness to pay for each attribute. A firm's offer function, Θ_i, for yogurt, for example, is determined by price, P_Y, product attributes, Z_i, and expected profit, Π_i: $\Theta_i=\Theta_i (P_Y, Z_i, \Pi_i)$. Different firms have comparative advantages in the production of different characteristics. Thus, firms offer milk or yogurt with different characteristics, that consumers want at different prices.[4]

The milk and yogurt markets are each assumed to be in equilibrium. Thus, where a firm's offer function is equal to a consumer's bid function, the marginal cost of production is equal to the marginal valuation of the consumer, which is the price. Differences among consumers in their desire for different milk and yogurt attributes and differences among firms in their capabilities of producing milk and yogurt with different characteristics leads to a variety in the types of these goods. The price of each product is a function of its characteristics. Characteristics include quantitative attributes such as the percentage of calcium and vitamin D and qualitative components such as if the product is 'organic' or 'all natural.'

III. Data and Variables

Data were collected from 31 grocery stores in suburban New Jersey.[5] This area was used because it contained all the grocery stores where a typical resident in the region could shop,

[3] I assume that the basic assumptions necessary for the hedonic model hold here. See Freeman (1993) for details on all of the necessary requirements.

[4] The same model holds for milk.

[5] The cities were: Edgewater, North Bergen, Fort Lee, Maplewood, Lodi, Passaic and Irvington.

though it is likely that a person would usually frequent only 2 or 3 different stores. Primary data was collected in the spring of 1999 by the author and graduate assistants. The grocery stores Pathmark, Shoprite, Kings, A & P and Grand Union were chosen because all natural, organic and non-organic goods were all available there. Convenience stores, such as 7-11, were excluded because they rarely, if ever, carry organic foods. The locations were also chosen because all of the stores were in upper-middle class areas where consumers would likely consider the choice of organic products.[6]

The stores provided 247 unique observations for milk and 768 unique observations for yogurt; that is there were 247 different brand sizes of milk total in each of the 31 stores and 768 types and brand sizes of yogurt total in the stores. Information about each dairy product's price and characteristics was recorded for both milk and yogurt. Tables 1a and 1b provide a description of the variables and give summary statistics for the milk and yogurt data respectively.

Milk is either organic or it is not. There is no label specifying that it is 'natural' or 'all natural'. Non-organic milk tends to be from local dairies in New Jersey, while organic milk is from another state. In addition, unlike yogurt, most milk does not have a national brand. Thus, milk producers either make organic milk or they do not, and none of the dairies in the data set sold both organic and non-organic milk.

There is a greater variety of product attributes for yogurt than for milk.[7] In addition to having different flavors, toppings and other characteristics, yogurt is also 'natural' or 'all natural' in addition to being organic or not. Some companies sell both milk and yogurt, usually those that provide organic products. Unlike milk, some yogurt manufacturers offer both organic and non-organic yogurt. Labels and brand names are much more important for yogurt than for milk. With the exception of store brands and some regional yogurt brands that compose a small percentage of the market, almost all yogurt is a nationally branded product.

Yogurt can be labeled organic, all natural or have no label indicating either of these. Thus, there are three levels of naturalness: organic, all natural and neither. Some firms specialize in one of these types of yogurt, some make two of the three and still others make all three types.

IV. Empirical Model and Results for Milk

A Cook-Weisberg test determined that variables were heteroskedastic using price as the dependent variable but not using price per ounce; thus price per ounce is used as the dependent variable.[8] Also due to severe multi-collinearity many of the variables had to be left out of the

[6] Regressions were estimated including dummy variables for the stores to control for possible differences between them. The results were similar to regressions estimated without the dummy variables. This is due to the fact that the stores were initially chosen because of their homogeneity. In order to preserve degrees of freedom final regressions discussed in the paper do not include dummy variables for the stores.

[7] Jell-O packs, Snackwell pudding and other non-yogurt products sold near yogurt were not included in the analysis.

[8] A Cook-Weisberg test showed that variables were heteroscedastic using price as the dependent variable but not using price per ounce.

regressions. For example, calories, fat calories, percent milk fat and other variables are all highly correlated, so only percent milk fat was included in the model. The functional form of the hedonic model is debatable. Since the primary purpose is to test whether organic milk sells for a premium over non-organic milk, a Box-Cox transformation, $P_h^{[\lambda]} = (P_h^{[\lambda]}-1)/\lambda$, that does not impose any functional form on the variables, was used. However, the coefficients of the independent variables cannot be interpreted in a Box-Cox regression.

Tests were conducted to see if linear or log linear regressions could be estimated. Different values of λ produce different functional forms: if λ equals one the functional form is linear, if λ equals zero the form is log linear. A Chi-squared test rejected a log linear and linear functional form at all levels of significance.[9]

Since larger containers are likely to sell at a discount relative to smaller containers, even after adjusting for price per ounce, dummy variables for container size are included. Perfect multi-collinearity among variables measuring dairy location occurred, because all of the non-organic milk was from New Jersey and all organic milk was from out of state so perfect multi-collinearity exists. Hence, these variables were excluded. Chocolate milk, lactaid-free milk, and other milk products without organic equivalents were also excluded from the analysis.[10]

The final specification of the estimated model estimates the price of milk per ounce as a function of the percent milk fat, protein, carbohydrate, calcium, vitamin D, vitamin A, 'organic' label, size of container and store.

Results of the regression are in Table 2a. As expected there is a statistically significant price premium for organic milk compared to its non-organic counterpart after controlling for other variables. Some consumers are willing to pay a premium for organic milk after controlling for other attributes. This premium is likely due to the desire to avoid consumption of the bovine growth hormone or avoid dairy products where cows were fed grain grown using herbicides and pesticides. While there is some debate whether organic food is healthier than its non-organic counterparts, consumers apparently believe that organic milk is more desirable. Since milk is not labeled as 'natural' or 'all natural,' there appears to be no confusion between the two different types of milk.

While the coefficients cannot be interpreted, the summary statistics show that the average price of a half-gallon of organic milk in the data set is $2.96 compared to the average price of a half-gallon of non-organic milk is $1.77. The higher price is due to the higher cost of production of not using hormones in organic milk and other items that reduce the cost of non-organic milk.

V. Empirical Model and Results for Yogurt

Again, because there is disagreement about which functional form should be used for a hedonic model, a Box Cox transformation that does not impose any restrictions on the form was

[9] The test statistic for $\lambda=0$ was $\chi^2(0) = 111.83$, and for $\lambda= 1$, $\chi^2(1)=328.34$.

[10] Some producers now offer organic chocolate milk but did not when the data was collected.

used for the first regression for yogurt. A Chi-squared test did not reject a linear functional form at any level of significance, but a log-linear functional form was rejected at all levels of significance. Thus, results using variables in a simple linear form are also presented in Table 2b and are discussed below.[11]

Multi-collinearity between many of the characteristics precipitated the exclusion of some variables. For example, fat calories, cholesterol and other variables related to calories were excluded. In addition, container size could not be included because of multi-collinearity with some brands of yogurt.

Since all natural and organic are mutually exclusive, yogurt labeled as organic was not also placed in the all natural category. However, there is some collinearity between some of the brand names and the all natural and organic variables. The correlation coefficient between Stonyfield and all natural was 0.63 and between Horizon and organic was 0.65.[12] All of the other correlation coefficients were under 0.3. Regressions were also estimated excluding the brand categories, with results similar to those reported below. I included brands in the final model to capture consumers' willingness to pay a premium for brand name yogurts.

To control for differences in sizes the dependent variable is price per ounce. The estimated model estimates the price of yogurt per ounce as a function of calories, protein, and dummy variables for topping outside the container, custard, fruit, all one flavor mixed together, flavor on the bottom, 'all natural' label, 'organic' label, brand of yogurt and the store.

Results are shown in Table 2b. All of the non-natural coefficients have the expected sign and most are statistically significant at the one percent level. Controlling for other yogurt characteristics, consumers are willing to pay a premium for brand name yogurt compared to the store brand. Some stores sell the yogurt at a higher price than others. Consumers are also willing to pay more for yogurt with fruit and a topping outside the container.

Compared to yogurt with no label proclaiming it to be 'all natural' or 'organic', the all-natural and organic coefficients are positive and statistically significant at the 1 percent level. However, given that the size of the 'all natural' and 'organic' coefficients were almost identical in both regressions and the t-statistics were also very similar it appears that there is no difference in consumer willingness to pay for organic yogurt and all natural yogurt.

To formally test this, an F-test (following Green 1993), of the form, $\beta_{\text{All Natural}} = \beta_{\text{Organic}}$, was performed to see if the all natural coefficient was significantly different from the organic coefficient. The test found that the two coefficients are not statistically significantly different at any level. Thus, there is no difference between the organic coefficient and the all-natural coefficient.

[11] The test statistic for $\lambda=0$ was $\chi^2(0) = 141.35$, and for $\lambda= 1$, $\chi^2(1)$ 0.8883.

[12] Stonyfield and Horizon both make all natural and organic yogurt. Other firms such as Breyers make both all natural and regular yogurt.

There are three main levels of 'naturalness' for yogurt: non-natural and non-organic, natural but not organic, and organic. The requirements for dairy products to be labeled 'organic' are more stringent than to meet requirements to be considered 'all natural'. Organic yogurt is seemingly healthier than its non-organic counterparts, since organic yogurt comes from cows that are fed grains free of pesticides or herbicides.

It would seem that consumers should be willing to pay more for each additional level of 'naturalness' but they don't.[13] It appears that firms have been able to take advantage of confusion over the terms 'all natural' and 'organic' and charge consumers more for the yogurt labeled 'all natural'. Consumers either think that the terms 'all natural' and 'organic' mean the same thing or do not care about the difference between them in the case of yogurt.

VI. Conclusion

Since consumers are willing to pay a significant premium for organic milk, it would seem that they would be willing to pay a premium for organic yogurt above 'all natural' yogurt. Yogurt that is 'all natural' can contain milk that was produced using BGH. The cows for 'all natural' milk could have also been fed food that contained genetically modified grains.

As technology plays an increasing role in the production of food, it is likely that consumers will want to know more about what is in their food. This will lead to the desire for more information. Knowing the difference between natural and organic will be necessary.

Policymakers in the Food and Drug Administration, the Department of Agriculture and elsewhere are going to have to grapple with the problem of how to regulate genetically altered food and certify organic food. There are also questions about what it means for food to be natural or organic. Perhaps the best policy is to simply provide information.
The USDA's new labeling requirements break organic foods into four categories in order to provide more information for consumers and lessen confusion. When making the announcement regarding the USDA's new standards, Secretary of Agriculture Dan Glickman said the new regulations create "a single national organic standard, backed by consistent and accurate labeling, that will greatly reduce consumer confusion." (Chicago Tribune, March 8, 2000)

The USDA's new regulations are a step in the right direction of providing consumers with more information. It may also be necessary to provide labeling that distinguishes what 'all natural' is and how it differs from the categories of organic.

[13] Other researchers have also found consumers' to be confused by labeling. For example, Morris, et.al. (1995) found the terms "recycled" and "recyclable" to be misunderstood.

References

Anstine, Jeff. "Consumers' Willingness to Pay for Recycled Content in Plastic Kitchen Garbage Bags: A Hedonic Price Approach," *Applied Economic Letters*, 2000, 7(1): 35-39.

Dhar, Tirtha and Jeremy Foltz. "Milk By Any Other Name… Consumer Benefits From Labeled Milk," *American Journal of Agricultural Economics*, 2005, 87(1): 214-228.

Freeman, A. Myrick. *The Measurement of Environmental and Resource Economics: Theory and Methods*, First Edition. 1993. Washington D.C.: Resources for the Future.

Green, William. *Econometric Analysis*, Second Edition. 1993. New York, NY: Macmillan Publishing Company.

Ippolito, P.M. and A.D. Mathios. "New Food Labeling Regulations and the Flow of Nutrition Information to Consumer," *Journal of Public Policy & Marketing* 1993, 12(2): 188-205.

Martin, Andrew. "Organic Milk Supply Expected to Surge as Farmers Pursue a Payoff," New York Times, April 20, 2007.

Morris, Louis, Manoj Hastak and Michael Mazis. "Consumer Comprehension of Environmental Advertising and Labeling Claims," *The Journal of Consumer Affairs*, 1995, 29(2): 328-350.

Nimon, Wesley and John Beghin. "Are Eco-Labels Valuable? Evidence From The Apparel Industry," *American Journal of Agricultural Economics*, 1999, 81(4): 801-811.

Organic Trade Association. www.ota.com

Park, Timothy and Luanne Lohr. "Supply and Demand Factors for Organic Produce," *American Journal of Agricultural Economics*, 1996, 78(3): 647-655.

Stanley, Linda and Stanley Tschirhart. "Hedonic Prices For A Nondurable Good: The Case Of Breakfast Cereals," *Review Of Economics And Statistics*, 1991, 73(3): 537-541.

Thompson, Gary. "Consumer Demand For Organic Foods: What We Know And What We Need To Know," *American Journal of Agricultural Economics*, 1998, 80(5): 1113 - 1118.

Thompson, Gary and Julia Kidwell. "Explaining The Choice Of Organic Produce: Cosmetic Defects, Prices And Consumer Preferences," *American Journal of Agricultural Economics*, 1998, 80(2): 277-287.

United States Department of Agriculture. 2005. "Organic Food Standards and Labels: The Facts" http://www.ams.usda.gov/nop/Consumers/brochure.html www.usda.gov.

West, Gale, Carole Gendron, Bruno Larue and Remy Lambert. "Consumers' Valuation of Functional Properties of Foods: Results from a Canada-wide Survey," *Canadian Journal of Agricultural Economics*, 2002, 50(4): 541-558.

Table 1a: Description of Variables for Milk and Summary Statistics

Variables	Description	Mean	Minimum	Maximum
Size	Size of container in ounces	65.6	32	128
Price	Price in dollars	2.14	.89	3.72
Calories	Number of calories per serving size, 1 cup (240 ml)	115.4	80	150
Fat calories	Number of calories from fat per serving size	31.7	0	70
Fat free	If label says 'Fat free', (yes=1)	.26	0	1
Low fat	If label says 'Low fat', (yes=1)	.28	0	1
Percent Milk fat	Percentage of Milk fat	1.4	0	3
Total fat	Total fat as a percentage of daily value	5.7	0	12
Saturated fat	Saturated fat as a percentage of daily value	11.4	0	25
Cholesterol	Cholesterol as a percentage of daily value	5.9	.5	20
Protein	Number of grams of protein	8.2	8	11
Sugars	Number of grams of sugars	12	10	16
Carbohydrate	Total carbohydrate as a percentage of daily value	4.6	4	12
Calcium	Calcium as a percentage of daily value	30	25	40
Sodium	Sodium as a percentage of daily value	5.2	1	7
Vitamin D	Vitamin D as a percentage of daily value	24.6	10	30
Vitamin A	Vitamin A as a percentage of daily value	9.3	0	15
Vitamin C	Vitamin C as a percentage of daily value	2.6	0	4
Pasteurized	If milk is pasteurized (yes=1)	1	1	1
Homogenized	If milk is homogenized (yes=1)	1	1	1
Organic	If milk is organic (yes=1)	.19	0	1
Gallon	If container size is one gallon (yes=1)	.31	0	1
Half Gallon	If container size is a half gallon (yes=1)	.42	0	1
Quart	If container size is a quart (yes=1)	.26	0	1
Grand Union	Store is Grand Union (yes=1)	.29	0	1
Kings	Store is Kings (yes=1)	.17	0	1
Shoprite	Store is Shoprite (yes=1)	.19	0	1
A & P	Store is A & P (yes=1)	.11	0	1
Pathmark	Store is Pathmark (yes=1)	.24	0	1

Table 1b: Description of Variables for Yogurt and Summary Statistics

Variable	Description	Mean	Minimum	Maximum
Size	Size of container in ounces	7.546875	6	8
Price	Price in dollars	.8488932	.40	1.27
Calories	Number of calories per serving size, 1 cup (240 ml)	178.8021	90	280
Fat calories	Number of calories from fat per serving size	13.65234	0	90
Fat free	If label says 'Fat free', (yes=1)	.4296875	0	1
Light	If label says 'Light', (yes=1)	.2434896	0	1
Artificial Sweetener	If yogurt contains aspartame (yes=1)	.2044271	0	1
Total fat	Total fat as a percentage of daily value	2.46224	0	15
Saturated fat	Saturated fat as a percentage of daily value	4.766927	0	30
Cholesterol	Cholesterol as a percentage of daily value	3.188802	0	13
Protein	Number of grams of protein	8.09375	5	15
Sugars	Number of grams of sugars	29.57813	10	72
Carbohydrate	Total carbohydrate as a percentage of daily value	10.9375	4	19
Calcium	Calcium as a percentage of daily value	28.6263	20	45
Sodium	Sodium as a percentage of daily value	5.106771	2	7
Potassium	Potassium as a percentage of daily value (Some containers did not give number)	10.63038	0	16
Topping	If the yogurt has a topping, such as nuts, separate from the yogurt	.0208333	0	1
Custard	If the yogurt is custard	.046875	0	1
Flavor on bottom	If the flavor has to be stirred from the bottom of the container	.3138021	0	1
Fruit	If the yogurt is fruit flavored	.5820313	0	1
All one flavor	The flavor is mixed in with the yogurt	.3216146	0	1
Plain	If the yogurt is plain	.0260417	0	1
All Natural	If the yogurt is labeled 'all natural'	.2200521	0	1
Organic	If yogurt is labeled 'organic' (yes=1)	.1041667	0	1
Dannon	The brand is Dannon	.2851563	0	1
Stonyfield	The brand is Stonyfield	.1028646	0	1
Breyers	The brand is Breyers	.1210938	0	1
Yoplait	The brand is Yoplait	.0572917	0	1
Columbo	The brand is Columbo	.0833333	0	1
LaYogurt	The brand is LaYogurt	.0885417	0	1
Yofarm	The brand is Yofarm	.0247396	0	1
Horizon	The brand is Horizon	.0520833	0	1
Brown Cow	The brand is Brown Cow	.0195313	0	1
Store Brand	The brand is Store Brand	.1653646	0	1
Grand Union	The Store is Grand Union	.2226563	0	1
Kings	The Store is Kings	.0390625	0	1
Shoprite	The Store is Shoprite	.3033854	0	1
A&P	The Store is A&P	.1653646	0	1
Pathmark	The Store is Pathmark	.1888021	0	1

Table 2a: Results for Milk, Dependent Variable Price Per Ounce of Milk

Independent Variables	Coefficient	T-statistic
Percent Milk fat	.8855462	1.540
Protein	7.909618	8.630***
Carbohydrate	.2671912	0.959
Calcium	-.7174839	-2.082**
Vitamin D	.3026935	1.336
Vitamin A	.390969	1.396
Organic	34.92518	25.341***
Gallon [a]	-42.54543	-33.898***
Half Gallon	-12.39429	-9.976***
Grand Union [b]	-1.377321	-0.931
Kings	-.1836171	-0.107
Shoprite	-4.791947	-3.076***
A & P	-11.81062	-6.152***
Intercept	-116.9269	

a: excluded size is quart
b: excluded store is Pathmark
* Significant at 10% level
** Significant at 5% level
*** Significant at 1% level

Number of observations: 247
F(13,233): 207.78
R-squared: 0.9206
Adjusted R-squared: 0.9162

Table 2b: Results for Yogurt: Dependent Variable: Price of Yogurt Per Ounce

Independent Variables	Box Cox Coefficient	Box Cox T-statistic	Linear Coefficient	Linear T-statistic
Calories (number per serving)	.0013098	2.838***	.002764	2.896***
Protein (number of grams)	-.0188398	-1.156	-.0405011	-1.202
Topping (outside yogurt, yes=1)a	.3616448	3.066***	.7306675	2.995***
Custard (yes=1)	-.4495192	-2.489**	-.9658323	-2.585**
Fruit (if flavor is fruit, yes=1)	.2328526	5.478***	.4570495	5.198***
All one flavor (yes=1)	-.0515449	-0.798	-.1088799	-0.815
Flavor on Bottom (yes=1)	-.2707694	-4.970***	-.5628499	-4.994***
'All Natural' Label (yes=1)	.3441454	4.124***	.6918978	4.008***
'Organic' Label (yes=1)	.3479923	3.586***	.7493074	3.733***
Dannon (yes=1)	2.843593	49.688***	5.541788	46.815***
Stonyfield (yes=1)	2.796331	23.932***	5.498708	22.752***
Breyers (yes=1)	2.383392	29.015***	4.597747	27.060***
Yoplait (yes=1)	3.115272	17.488***	6.175193	16.760***
Columbo (yes=1)	2.547898	24.804***	4.936149	23.232***
LaYogurt (yes=1)	1.791129	17.088***	3.399224	15.678***
Yofarm (yes=1)	1.743699	12.672***	3.301114	11.598***
Horizon (yes=1)	4.901185	35.713***	10.17103	35.830***
Brown Cow (yes=1)	4.660703	28.007***	9.703639	28.191***
A&P (yes=1)	-.1591049	-2.909***	-.2926806	-2.587**
Grand Union (yes=1)b	.0624336	1.244	.217278	2.094**
Shoprite (yes=1)	-.0784755	-1.632	-.1334628	-1.341
Kings (yes=1)	-.1128654	-1.908*	-.2025867	-1.656
Intercept	3.864865	26.598	6.438838	21.423
* Significant at 10%	Number of Observations: 768		Number of Observations: 768	
** Significant at 5%	F (23, 744): 298.01		F (23, 744): 289.66	
*** Significant at 1%	R-squared: 0.9021		R-squared: 0.8995	
a: plain is excluded	Adjusted R-squared: 0.8991		Adjusted R-squared: 0.8964	
b: Pathmark is excluded				

An Analysis of Health Care Utilization Controlling for Selectivity of Health Plan Choice

Christina H. Rennhoff*

Abstract

Due to rising health care costs, it is increasingly important that we understand the health care utilization decisions of individuals. In this paper, we use the personnel and health claim data of a large employer to examine the relationship between health care utilization and worker characteristics. Due to the possibility of endogenous health plan choices, simple OLS health care utilization regressions are potentially biased and would provide inconsistent parameter estimates. We estimate a model of health care utilization that addresses the problem of selectivity of health plan choice. We find evidence of self selection when the utilization equation includes individual characteristics of the workers.

I. Introduction

Given the recent attention paid to health care costs, it is increasingly important that we understand the health care utilization decisions of individuals. Using data on health care consumption and personal characteristics of individuals, we can use econometric techniques to estimate a utilization equation that will help us understand the relationship between individual characteristics and health care utilization. In order to accurately estimate a health care utilization equation, it is necessary to correct for possible selection bias caused by health plan choice. This is because the sample used to estimate the health care utilization equation for any health plan consists only of individuals who have selected that particular health plan. Since health plan selection is non-random and unobserved individual characteristics affecting health plan choice may also influence health care utilization, the possibility of selection bias arises. For example, suppose an individual's health status is unobserved. If an individual has poor health status, she may utilize a lot of health care and select a health plan with generous benefits. Therefore, this unobserved characteristic affects both health plan choice and health care utilization. When selection bias exists, simple OLS health care utilization regressions, which treat health plan choices as exogenous, are biased and provide inconsistent parameter estimates. In order to estimate consistent parameters, it is necessary to treat health plan choices as endogenous.

Using the health claim and personnel dataset from a large employer, this study estimates a health care utilization equation and reconciles the selection problem by applying the selection correction technique proposed by Lee (1983).[1] Lee's technique is a two step approach that allows

*St. Joseph's University, Assistant Professor, Department of Economics, 5600 City Avenue, Philadelphia, PA 19131. E-mail: christina.rennhoff@sju.edu

[1] Ninety percent of privately insured individuals obtain their health insurance through an employer (their own, their spouse's or their parents'). (Gruber, 2001)

for selection correction and is applied in the following empirical studies on sample selection in polychotomous choice models: Trost and Lee (1984), Gyourko and Tracy (1988), Johnes (1999), Harris (1993), Dowd et al. (1991) and Zhang (2004).

Previous studies that estimate health care utilization equations controlling for the selectivity of health plan choice include Cameron, et al. (1988), Savage and Wright (2003) and Dowd, et al. (1991). The primary advantage of this study over the previous studies is the dataset used in this study. All three previous studies rely upon survey data. Surveys depend on the ability of respondents to recall their past behavior accurately. Given that respondents may not accurately remember past behaviors and they suffer no consequence from reporting misinformation, survey data may possibly be less accurate than health claim data. These studies also suffer from one or more of the following: limited information on health care utilization, missing information on the prices of health care and health insurance, or health care utilization information over a very short time period. In contrast, the dataset used in this study is very accurate, spans a one year time period and includes detailed information on worker characteristics, health care utilization and health plan choice.

The remainder of this paper is organized as follows. Section II discusses the data used in estimation and Section III introduces the model. Results from estimation are discussed in Section IV and Section V concludes.

II. Data

Undertaking a study of this nature requires health plan choice and health care utilization information for a group of individuals. Fortunately, the State of Tennessee (hereafter SOT) has granted us access to the health claim and personnel data of their workers for the purposes of this study.[2] The data span from January to December 2001. The personnel data includes information regarding worker characteristics such as age, race, sex, counties in which the worker lives and works, salary and health plan choice. With the exception of pharmaceutical claims, the SOT has also provided access to all health claims for workers during the calendar year of 2001. The health claim data includes the date, the diagnosis codes (ICD 9 codes) and the associated costs (co-pay, deductible and net payments made by the SOT for each claim).

The SOT offers three different health plans to their workers, a Preferred Provider Organization plan (PPO), a Health Maintenance Organization plan (HMO) and a Point of Service plan (POS). A benefit summary of the three plans is presented in Table 1. The PPO plan charges the highest premium and, in some respects, offers the most extensive benefits. Enrollees in the PPO plan have the option of seeking care through an in-network provider or an out-of-network provider (paying a higher co-pay for health care administered by an out-of-network provider). In addition, enrollees in the PPO plan have an annual out-of-pocket maximum. Any health care received after the out-of-pocket maximum is met at no cost to the enrollee.

The HMO charges the lowest premium of the three plans and can be thought of as offering the least extensive benefits. Only health care sought by an in-network provider is covered by the

[2] The primary source of the data was provided in such a way that the subjects cannot be identified, directly or through identifiers linked to the subjects.

HMO plan and there is no annual out-of-pocket maximum for the enrollees. The POS plan is somewhat in the middle. The premium for the POS plan is lower than that of the PPO, and higher than that of the HMO. Enrollees in the POS plan have the option of seeking care through an out-of-network provider (paying a higher co-pay), but there is no annual out-of-pocket maximum for their expenses. Additionally, POS and HMO enrollees must have a referral from their primary care physician in order for health care provided by a specialist to be covered by their health plan.

For the purposes of this paper, a number of deletions were made to the original dataset. The original dataset was comprised of 41,665 employees. First, all employees who were not employed by the SOT for the entire year were deleted from the sample. After imposing this restriction, the sample shrank from 41,665 employees to a sample of 34,633. Secondly, all employees with missing health plan choice information were deleted from the sample. These employees were either uninsured or purchased health insurance through another provider (perhaps a spouse's employer). Since we have no information regarding the health insurance status or health care utilization of these employees, we would have difficulties accurately modeling their health care decisions. After deleting these employees, the sample was then comprised of 31,771 employees. Third, we deleted all married couples who were both employed by the SOT, because they had different health plan options than the other employees. This deletion further reduced the sample to 29,522 employees. The fourth deletion was to exclude all employees who were younger than 17 years old or older than 88 years old. This fourth deletion shrank the sample from 29,522 to 29,503 employees. The fifth deletion excluded all employees who were employed in a county that did not offer all three health plans. This deletion shrank the sample from 29,503 to 24,527 employees. Finally, we only included employees that were enrolled in single plans (rather than family plans). We examine single plan enrollees because in the case of family plan enrollees the reported demographic and socioeconomic information pertains to the worker who is not necessarily the patient receiving the health care. When the spouse or child of a family plan enrollee seeks health care, the demographic and socioeconomic information would not be consistent with the person seeking care. In the case of single plan enrollees, however, the demographic and socioeconomic information would always pertain to the person seeking care. Single plan enrollees are not necessarily employees that are not married; they are simply employees who opted for health insurance coverage for themselves only and not their spouse or dependent(s) (in the case that they have spouses and/or dependent(s)). Once the deletions were made, we were left with 10,389 employees in our sample.

Table 2 shows the summary statistics of health plan enrollees by age and Table 3 shows the summary statistics of health plan enrollees by salary. Over 89% of workers enroll in either the PPO plan (which is the most comprehensive plan) or the HMO plan (which is the least comprehensive plan). Workers younger than age 50 tend to choose the HMO plan while older workers tend to choose the PPO plan (as shown by Table 2). In addition, the enrollees in the HMO plan tend to have lower salaries than PPO enrollees (as shown by Table 3). These patterns suggest that individuals select plans in ways that are endogenous with their observable characteristics which would bias estimates of the impact of such characteristics on health care utilization.

In this study, we use health care expenditures to represent health care utilization. Health care expenditures represent the total payment for health care services rendered (costs paid by enrollee plus costs paid by the insurer). Tables 2 and 3 also show the average health care

expenditures for enrollees by age and salary, respectively. Generally speaking, PPO enrollees utilize more health care than those enrolled in the HMO and POS plans. In addition, Table 2 indicates that older employees utilize more health care than younger employees. Table 3 suggests no clear relationship between salary and health care utilization.

III. Model

As noted in Section I, in order to accurately estimate a health care utilization equation, one must correct for the sample selection bias. To reconcile the problem of selection bias, we adopt Lee's (1983) estimation framework for modeling polychotomous choice problems with mixed continuous and discrete dependent variables. The model can be defined by the following two equations:

$$U_{ij} = z_i'\gamma_j + p_j'\delta + \mu_{ij} \tag{1}$$

$$\ln h_{ij} = x_i'\beta + \varepsilon_{ij} \tag{2}$$

The subscript i indexes workers (i =1, 2, ...n). The subscript j indexes health plans. Since workers are offered a choice of three health plans, j =1, 2, 3. The utility function of worker i, conditional on choosing health plan j, is defined in (1) and is used to estimate the discrete health plan choice of worker i. The natural log of the utilization of health care of worker i, conditional on choosing health plan j, is defined in (2) and is used to estimate the continuous health care utilization decisions of employees.[3] Worker characteristics and plan characteristics that affect plan choice are represented by z_i and p_j, respectively. Worker characteristics that affect health care utilization are represented by x_i. The two error terms μ_{ij} and ε_{ij} represent unobserved variables that affect utility and health care utilization, respectively.

The i^{th} worker is assumed to choose plan j if

$$U_{ij} > \max U_{ik} \quad k \neq j \text{ or}$$

$$z_i'\gamma_j + p_j'\delta + \mu_{ij} > \max U_{ik} \quad k \neq j \text{ or}$$

$$z_i'\gamma_j + p_j'\delta > e_{ij} \tag{3}$$

where

$$e_{ij} = \max U_{ik} - \mu_{ij} \tag{4}$$

[3] The natural log of health care utilization is used in (2). Therefore, in order to avoid dropping individuals from the sample that utilized zero amounts of health care, we replaced the zero amount of health care utilization with 0.00001.

Since we assume that e_{ij} is *iid* Gumbel distributed, the probability that worker *i* will choose plan j is

$$\Pr(e_{ij} < z_i'\gamma_j + p_j'\delta) = \frac{\exp(z_i'\gamma_j + p_j'\delta)}{\sum_{j=1}^{3}\exp(z_i'\gamma_j + p_j'\delta)} \tag{5}$$

The worker's health plan choice is, therefore, analyzed with a multinomial logit model. Using only workers that select into each health plan, the expected utilization of health care conditional on enrollment in the j^{th} health plan is

$$\begin{aligned} E \mid \ln h_{ij} \mid i \text{ chooses plan } j \mid &= E[\ln h_{ij} \mid e_{ij} < z_i'\gamma_j + p_j'\delta] \\ &= x_i'\beta + E[\varepsilon_{ij} \mid e_{ij} < z_i'\gamma_j + p_j'\delta] \end{aligned} \tag{6}$$

Because of the selection bias in the observed data, $E[\varepsilon_{ij} \mid e_{ij} < z_i'\gamma_j + p_j'\delta] \neq 0$. Therefore, the least squares estimation produces inconsistent estimates of β. In following Lee (1983), e_{ij} is transformed into a standard normal variable e_{ij}^* by

$$e_{ij}^* = \Phi^{-1}[F(z_i'\gamma_j + p_j'\delta] \tag{7}$$

where Φ^{-1} is the inverse standard normal CDF. Further,

$$e_{ij} < z_i'\gamma_j + p_j'\delta \text{ iff } e_{ij}^* < \Phi^{-1}[F(z_i'\gamma_j + p_j'\delta] \tag{8}$$

Substituting from (8) into the conditional expectation term in (6) yields

$$E[\ln h_{ij} \mid i \text{ chooses plan } j] = x_i'\beta + E[\varepsilon_{ij} \mid e_{ij}^* < \Phi^{-1}[F(z_i'\gamma_j + p_j'\delta)]. \tag{9}$$

Therefore, the conditional health care utilization can be evaluated using standard methods such that

$$E[\ln h_{ij} \mid i \text{ chooses plan } j] = x_i'\beta - \sigma_j\rho_j\left[\frac{\phi\{\Phi^{-1}[F(z_i'\gamma_j + p_j'\delta)]\}}{F(z_i'\gamma_j + p_j'\delta)}\right] \tag{10}$$

where ρ_j is the correlation coefficient between ε_{ij} and e_{ij}^*.

Consistent estimates for the βs can be obtained by replacing γ_j and δ with the first stage logistic estimates $\hat{\gamma}_j$ and $\hat{\delta}$ from (5) and estimating (10) using OLS. This substitution implies that the standard errors for the βs reported by OLS are biased since they assume that the first stage

logistic estimates are observed. The corrected covariance matrices are derived following the method discussed in Lee (1978).

IV. Results

In estimating the multinomial logit model of plan choice defined by (5), we included variables that were believed to be important in an individual's decision of which health plan to choose. These variables included the health plan premium and worker characteristics (such as age, gender, salary, race, marital status and a variable that indicates whether the worker lives or works in a metropolitan statistical area). The inclusion of the health plan premium helps us to identify the plan choice equation from the health care utilization equation that we will estimate in the second step. The results of the multinomial logit model of plan choice are reported in Table 4. For identification purposes, the parameters for worker characteristics for the POS plan are normalized to zero.

According to the results in Table 4, the coefficients for age are positive and significant for both the HMO and PPO plans and the coefficient for age squared is negative for both the HMO and PPO plans, but significant for the HMO plan only. This implies that as workers grow older, workers are more likely to choose the HMO and PPO plans over the POS plan, but due to the negative coefficient for age squared, this effect diminishes as the worker grows older.

The coefficients for female are negative for both the HMO and PPO plans, but are only significant for the HMO plan. Therefore, females are less likely than males to choose the HMO plan than the POS plan. The results also indicate that white workers and workers with higher salaries are more likely to choose the PPO plan over the POS plan and are less likely to choose the HMO plan over the POS plan. Given that the PPO plan has the highest premium and the most comprehensive coverage while the HMO has the lowest premium and the least comprehensive coverage, the results indicate that workers are more likely to enroll in the more comprehensive health plans that have higher premiums as their salaries rise. In addition, white workers are more likely than non-white workers to enroll in the more comprehensive health plans that have higher premiums.

The results also indicate that workers who live or work within a metropolitan statistical area are more likely to choose the HMO and PPO plans over the POS plan and married individuals are less likely to choose the PPO plan than the POS plan.[4] In addition, the negative and significant coefficient for health plan premium implies that as the premium rises, employees are less likely to choose that health plan.

The health care utilization equation estimates are given in Table 5 where both OLS and two stage selectivity corrected results are presented. We use age, salary and a set of dummy variables on gender, race, marital status, and whether the worker lives or works in a small or large

[4] Due to the nature of the data only workers enrolled in single plans (as opposed to family plans) are included in the study. Therefore, in this study a worker that is married is enrolled in a single health plan that only covers the worker and not his/her spouse or dependents. These married individuals may not be representative of the married population as a whole.

metropolitan area as independent variables. The dependent variable, which is used to represent health care utilization, is the natural log of total health care expenditures. The standard errors of the two stage estimates have been corrected to account for the fact that the first stage logistic estimates are not observed. Our discussion in this section is based on the two stage results.

According to the results in Table 5, the coefficient for age is positive and significant. Therefore, as individuals age, their health care utilization increases. Previous research that has estimated the effects of age and age squared on health care utilization have shown a U-shaped pattern of utilization: Dowd et al. (1991) and Cameron et al. (1988). A negative coefficient for age and a positive coefficient for age squared would exhibit this U-shaped pattern. However, when both age and age squared were included in this analysis, neither coefficient was significant. We, therefore, omit the age squared variable and focus solely on age.[5]

The coefficient for female is positive and significant. Therefore, according to the results, females utilize more health care than males on average. This result is not surprising given the previous evidence showing that females utilize more care than men even after controlling for gynecological and obstetrical care and for severity of medical problem (Sindelar, 1982). In addition, according to the results in Table 5, whites tend to utilize more health care than non-whites and married workers tend to utilize more health care than non-married workers.[6]

The variable metrobig is a dummy variable that indicates if the worker lives or works in a metropolitan statistical area that has a population of 500,000 or more. The variable metrosmall is a dummy variable that indicates if the worker lives or works in a metropolitan statistical area that has a population of less than 500,000. Workers whose dummy variables are zero for both metrobig and metrosmall live and work in a rural area. Given the fact that there tends to be more health care facilities in metropolitan areas versus rural areas, one would expect the coefficients for metrobig and metrosmall to be positive. Surprisingly, however, the coefficients for metrobig and metrosmall are negative but not significant.

The extent of worker self selection into each of the health plans is indicated by the coefficients on the select variables which are represented by the term $\frac{\phi\{\Phi^{-1}[F(z_i'\gamma_j + p_j'\delta)]\}}{F(z_i'\gamma_j + p_j'\delta)}$ indicated in (10). The coefficients of the selectivity variables are significant for the HMO and PPO plans. Therefore, for these two health plans, evidence of self selection is found.

It is possible to use the coefficients for the selectivity variable and estimate the difference between the amount of health care an individual who self selects into a particular health plan would

[5] One would expect health care utilization to be a function of an individual's salary. However, the coefficient for salary is not significant. As noted previously, the data includes individuals that are married and not married. Therefore, while the salary information may be a good proxy for household income for non-married individuals, the household income for married individuals may also include a spouse's salary. Unfortunately, we do not observe this information. Because of this, the salary variable may be measured with noise.

[6] Since we only examined single plan enrollees, married individuals are individuals who opt for single health coverage that only covers themselves and not their spouse or dependent(s). These married individuals may not be representative of the married population as a whole.

utilize and the amount of health care a randomly drawn individual with identical characteristics would utilize under that plan.[7] Since the coefficients are significant for the HMO and PPO plans, we present these estimates for the HMO and PPO plans only. The estimates indicate that workers who select the HMO plan utilize on average 11.3% more health care than an "identical" randomly drawn worker would utilize under the HMO plan and that workers who select the PPO plan utilize on average 9.7% more health care than an "identical" randomly drawn worker would utilize under the PPO plan.

V. Conclusion

Using data of a group of workers that chose between three employer provided health plans, a general selection model was used to estimate health care utilization equations. The error terms in health care utilization equations are often truncated by the endogenous choice of health plans. The model in this paper provides an approach to correcting this problem. When the utilization equation includes individual characteristics of the workers, we find evidence of self selection in two of the three health plans. The results from this paper indicate that controlling for coverage, individuals who selected the HMO and PPO plans would utilize different amounts of health care if they selected alternative health plans.

[7] These estimates are calculated by multiplying the selection coefficient (-[$\sigma_j\rho_j$]), times the mean value of the selection variable for workers who selected that health plan. See (10) in text. (Gyourko & Tracy, 1988)

References

Cameron, A., P. Trivedi, F. Milne, and J. A. Piggott. "Microeconometric Model of the Demand for Health Care and Health Insurance in Australia." *The Review of Economic Studies* 1988; 55: 85-106.

Dowd, B., F. Roger, S. Cassou, and M. Finch. "Health Plan Choice and the Utilization of Health Care Services." *The Review of Economics and Statistics* 1991; 73: 85-93.

Gruber, J. "Taxes and Health Insurance." NBER Working Paper #8657 2001.

Gyourko, J., and J. Tracy. "An Analysis of Public and Private Sector Wages Allowing for Endogenous Choices of Both Government and Union Status." *Journal of Labor Economics* 1988; 6: 229-53.

Harris, R. "Part-time Female Earnings: An Analysis Using Northern Ireland NES Data." *Applied Economics* 1993; 25:1-12.

Johnes, G. "Schooling, Fertility and the Labour Market Experience of Married Women." *Applied Economics* 1999; 31:585-92.

Lee, L. F. "Generalized Econometric Models with Selectivity." *Econometrica* 1983; 51: 507-12.

Lee, L. F. "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables." International Economic Review 1978; 19:415-33.

Savage, E. and D. Wright. "Moral Hazard and Adverse Selection in Australian Private Hospitals: 1989-1990." *Journal of Health Economics* 2003; 22: 331-59.

Sindelar, J. "Differential Use of Medical Care by Sex." *Journal of Political Economy* 1982; 90: 1003-19.

Trost, R. P. and L. F. Lee. "Technical Training and Earnings: A Polychotomous Choice Model with Selectivity." *The Review of Economics and Statistics* 1984; 66:151-6.

Zhang, H. "Self Selection and Wage Differentials in Urban China: A Polychotomous Model with Selectivity." Unpublished manuscript. 2004.

Table 1
Overview of Plan Characteristics
Summary of Benefits

	HMO	POS	PPO
Premium Family/Single			
Family	$702.66	$1,014.96	$1,415.52
Single	$272.44	$406.44	$566.88
Co-pay (Physical Care)			
In-network†			
PCP	$10	$15	10%
Specialist	$15	$15	10%
Out-of-network			
PCP	NONE	30%	30%
Specialist	NONE	30%	30%
Co-pay (Mental Care)			
Outpatient	$15 / 45 visits	$15 / 45 visits	$5 (1-15 visits) $25 (16-45 visits)
Inpatient††	$100 / 30 days	$100 / 30 days	10% / 45 days
Employee Assistance*	NO	YES	YES
PCP Referral for Specialist	YES	YES	NO
Partially cover care from out-of-network providers	NO	YES	YES
Deductible	NO	NO	YES
Annual out-of-pocket max**	NO	NO	YES
Assignment payments***	YES	YES (in-network) NO (out-of network)	NO

†The 10% co-pay for in-network care under the PPO plan is generally more expensive than the $15 copay under the POS plan and the $10 and $15 copays under the HMO plan.
††The 10% co-pay for inpatient mental health care under the PPO plan is generally more expensive than the $100 copay under the POS and HMO plans.
*Employee Assistance Program is a type of managed care behavioral health (carve-out plan). The insured employee seeks help for mental health care by calling a specialist who will refer the patient to the appropriate mental health provider.
**Excludes mental health benefits.
***Assignment is a form of health payment. Under assignment, health care providers send the bill directly to the insurer, and the patient pays a co-pay. If the form of payment is not assignment, then it is individual reimbursement. Under individual reimbursement, the patient pays all the charges, sends copies of the bills to the insurer, and is reimbursed.

Table 2
Summary Statistics by Age

Age	HMO	POS	PPO	Total
18-30				
Enrollment	921	229	340	1,490
%	61.81%	15.37%	22.82%	100.00%
Average health care expenditures*	$1.4	$1.1	$1.6	$1.4
31-40				
Enrollment	893	177	557	1,627
%	54.89%	10.88%	34.23%	100.00%
Average health care expenditures*	$1.5	$1.7	$1.6	$1.5
41-50				
Enrollment	1,433	332	1,351	3,116
%	45.99%	10.65%	43.36%	100.00%
Average health care expenditures*	$2.2	$2.3	$2.4	$2.3
51-60				
Enrollment	1,195	306	1,751	3,252
%	36.75%	9.41%	53.84%	100.00%
Average health care expenditures*	$2.7	$3.0	$2.8	$2.8
over 60				
Enrollment	238	60	606	904
%	26.33%	6.64%	67.04%	100.00%
Average health care expenditures*	$4.0	$3.4	$4.1	$4.0
Total				
Enrollment	4,680	1,104	4,605	10,389
%	45.05%	10.63%	44.33%	100.00%
Average health care expenditures*	$2.1	$2.2	$2.6	$2.3

* Average health care expenditures are represented in thousands of dollars.

Table 3
Summary Statistics by Salary

Salary	HMO	POS	PPO	Total
< $20,000				
Enrollment	977	151	435	1,563
%	62.51%	9.66%	27.83%	100.00%
Average health care expenditures*	$2.0	$1.7	$2.8	$2.2
$20,000-$40,000				
Enrollment	3,011	726	2,963	6,700
%	44.94%	10.84%	44.22%	100.00%
Average health care expenditures*	$2.2	$2.4	$2.6	$2.4
$40,000-$60,000				
Enrollment	617	184	991	1,792
%	34.43%	10.27%	55.30%	100.00%
Average health care expenditures*	$2.2	$2.2	$2.4	$2.3
> $60,000				
Enrollment	75	43	216	334
%	22.46%	12.87%	64.67%	100.00%
Average health care expenditures*	$1.7	$1.5	$2.9	$2.4
Total				
Enrollment	4,680	1,104	4,605	10,389
%	45.05%	10.63%	44.33%	100.00%
Average health care expenditures*	$2.1	$2.2	$2.6	$2.3

* Average health care expenditures are represented in thousands of dollars.

Table 4
Multinomial Logit Model of Health Plan Choice

	HMO	POS[†]	PPO
Age	0.0484**		0.0551**
	(0.0092)		(0.0094)
Age Squared	-0.0005**		-0.0001
	(0.0001)		(0.0001)
Female	-0.2162**		-0.0859
	(0.0705)		(0.0705)
Metro[††]	1.0686**		0.2177**
	(0.0889)		(0.0843)
Salary[†††]	-0.0256**		0.0047*
	(0.0030)		(0.0027)
White	-0.6146**		0.1503*
	(0.0818)		(0.0855)
Married	-0.0944		-0.2059**
	(0.0685)		(0.0684)
Premium	-7.0511**	-7.0511**	-7.0511**
	(1.0178)	(1.0178)	(1.0178)

[†]The parameters for the POS plan for worker characteristics (age, age squared, female, metro, salary, white and married) are normalized to zero for identification purposes.

[††]A dummy variable that indicates whether the worker lives or works in a metropolitan statistical area.
[†††]Salary is expressed in thousands of dollars.
* Significant at the 10% level.
** Significant at the 5% level.

Table 5
Health Care Utilization Equation Estimates
Dependent variable =
Natural log of total health care expenditures

	OLS	Two-Stage
Intercept	1.7524**	2.0101**
	(0.1970)	(0.2343)
age	0.0411**	0.0416**
	(0.0034)	(0.0035)
female	1.9691**	1.9641**
	(0.0806)	(0.0829)
salary	0.0012	0.0008
	(0.0031)	(0.0033)
white	0.5249**	0.5535**
	(0.0916)	(0.0952)
marry	0.2824**	0.2869**
	(0.0789)	(0.0813)
metrobig	-0.0555	-0.0385
	(0.1063)	(0.1101)
metrosmall	-0.3613*	-0.3285
	(0.2080)	(0.2142)
select hmo		-0.3839**
		(0.1618)
select ppo		-0.4821**
		(0.1609)
select pos		0.0407
		(0.1081)
R^2	0.0770	0.0788

* Significant at the 10% level.
** Significant at the 5% level.

Journal of Applied Economics and Policy

Call for Papers

The *Journal of Applied Economics and Policy* is an electronic journal published by the Kentucky Economic Association. The *Journal* will consider manuscripts in the following four categories:

1. Theory and Practice of Economics: The *Journal* will consider manuscripts from all JEL categories, but some preference will be given to manuscripts examining issues and policies relevant to Kentucky's economy and its socio-political and economic institutions.

2. Teaching of Economics: The *Journal* will consider manuscripts about teaching methods and empirical studies of teaching methodologies.

3. Student Papers: The *Journal* will consider manuscripts from student authors from all JEL categories.

4. Book Reviews: The *Journal* will consider reviews of books from any JEL category. Prior to submitting a book review, authors must contact the editor – Cathy.Carey@wku.edu – for additional information about book reviews.

Manuscripts should be sent to Cathy Carey, Department of Economics, Western Kentucky University, 1906 College Height Boulevard, Bowling Green, KY 42101. The submission fee is $20.00. Make all checks payable to *Journal of Applied Economics and Policy*. There is no submission fee for book reviews.