**Demographic and Economic Profiling of U.S. Demand for Probiotics:**

**The Case of Drinkable Yogurt**

Senarath Dharmasena*

Oral Capps, Jr.*

*Agribusiness, Food and Consumer Economics Research Center (AFCERC)

Texas A&M University

Emails: sdharmasena@tamu.edu

ocapps@tamu.edu

Phone: (979) 845-8491

# Demographic and Economic Profiling of U.S. Demand for Probiotics: The Case of Drinkable Yogurt

Senarath Dharmasena and Oral Capps, Jr.

## Abstract

Data from U.S. households for calendar year 2008 were used in examining demographic and economic factors affecting demand for drinkable yogurts using Heckman two-step procedure. Price, region, race and gender of household head were significant drivers of consumption of drinkable yogurts. Statistically significant sample selection bias was observed.

Key Words: Probiotics, drinkable yogurts, Nielsen HomeScan data, Heckman two-step

*JEL Classification: D11, D12*

**Demographic and Economic Profiling of U.S. Demand for Probiotics:**

**The Case of Drinkable Yogurt**

Senarath Dharmasena and Oral Capps, Jr.

**Background:**

There are many different types of non-alcoholic beverages available in the United States today compared to say the past decade. Functionality and health dimensions of beverages such as nutrient fortification, presence of beneficial cultured microorganisms, and enrichment with favorable fatty acids like Omaga-3, are important considerations along with conventional economic factors (Ozer and Kirmaci, 2010; Beverage Marketing Corporation, 2008). Probiotics, beverages with beneficial cultured microorganisms, are the main bioactive product of the functional dairy food category and are still in the forefront of innovation (Ozer and Kirmaci, 2010; Champagne, 2009). At present, probiotics is a $450 million market with roughly a 16% annual growth (Convenient Store News, 2010). More specifically, drinkable yogurt, defined as a dairy-based functional beverage that may or may not include fruit or fruit flavoring, is a probiotic non-alcoholic beverage which is relatively new and growing in the United States (Thompson *et al*., 2007).

Knowledge of price sensitivity, substitutes/complements and demographic profiling in particular is important for manufacturers, retailers and advertisers of drinkable yogurts from a competitive intelligence standpoint and from the standpoint of strategic decision making. We could find two past studies pertaining to drinkable yogurts in the extant literature. Thompson *et al*., 2007 investigated commercial strawberry drinkable yogurts linking sensory product characteristics to consumer preferences for Caucasian, African

American and Hispanic consumers in the United States. They observed variability in consumer acceptability across the three ethnic groups. Key drivers for all three groups were found to be natural strawberry flavor and sweet taste. Also, they found that drinkable yogurts with specific flavor and physical properties could be marketed to specific target market segments.

Davis *et al*, (2010) estimated own-price and cross-price elasticities for thirteen dairy product categories including drinkable yogurts using Nielsen HomeScan data for calendar year 2007. They used a censored almost ideal demand system model to come up with their estimates. The estimated compensated own-price elasticity for drinkable yogurt was -1.72. Age of the household head, presence of children, ethnicity and region were found to be significant demographic characteristics affecting the demand for drinkable yogurt.

**Objectives**

A thorough and a complete analysis of demand for drinkable yogurts is important primarily due to increasing growth in consumption in recent times and to the lack of information in the literature. In this light, specific objectives are: (1) to determine the factors affecting the decision to purchase drinkable yogurt, and (2) once the decision to purchase a drinkable yogurt is made, to determine the drivers of purchase volume.

**Methodology**

Initially, household purchases of drinkable yogurts (expenditure and quantity) and socio-economic-demographic characteristics are generated for each household in the Nielsen HomeScan Panel for calendar year 2008. As such we have total of 61,440 households, out of which only 11,546 unique households purchased drinkable yogurts. Using this data set, we

estimate demand for drinkable yogurts with adjustment to sample selection bias (Heckman, 1979).

Quantity data are standardized in terms of liquid ounces and expenditure data are expressed in terms of dollars. Then taking the ratio of expenditure to volume, we generate unit values (prices in dollars per ounce).

Factors hypothesized to affect the decision to buy drinkable yogurt and volume of drinkable yogurt purchased are: price of drinkable yogurt, income of the household head, and other demographic characteristics such as age, gender, employment and education status of the household head; region; race; Hispanic origin; age and presence of children.

**Model Development, Procedures and Variables**

Choice to purchase or not to purchase drinkable yogurt could be affected by price of drinkable yogurt and various demographic factors. This type of choice is a dichotomous discrete (buy or not-to buy or "one" if buy and "zero" if do not buy) and a probit model is used generally to model such a choice decision. The dependent variable is a zero one type dummy variable which is created to reflect the non-purchase or purchase respectively of drinkable yogurt. It is regressed on price and a host of demographic factors. Probit analysis will provide statistically significant findings of the decision to purchase drinkable yogurt.

Demographic and economic factors hypothesized to be affecting the decision to buy drinkable yogurt are listed on Table 1. Also, we provide different categories used in each factor along with base category for dummy variables.

The probit model for drinkable yogurt can be written as follows:

$$\Pr(Y = 1 \mid x_i'\beta) = \beta_1 + \beta_2 PRICE_i + \beta_3 AGEHH2529_i + \beta_4 AGEHH3034_i +$$
$$\beta_5 AGEHH3544_i + \beta_6 AGEHH4554_i + \beta_7 AGEHH5564_i + \beta_8 AGEHHGT64_i +$$
$$\beta_9 EMPHHPT_i + \beta_{10} EMPHHFT_i + \beta_{11} EDUHHHS_i + \beta_{12} EDUHHU_i +$$
$$\beta_{13} EDUHHPC_i + \beta_{14} MIDWEST_i + \beta_{15} SOUTH_i +$$
$$\beta_{16} WEST_i + \beta_{17} BLACK_i + \beta_{18} ASIAN_i +$$
$$\beta_{19} OTHER_i + \beta_{20} HISP\_YES_i + \beta_{21} AGEPCLT6\_ONLY_i +$$
$$\beta_{22} AGEPC6\_12ONLY_i + \beta_{23} AGEPC13\_17ONLY_i +$$
$$\beta_{24} AGEPCLT6\_6\_12ONLY_i + \beta_{25} AGEPCLT6\_13\_17ONLY_i +$$
$$\beta_{26} AGEPC6\_12AND13\_17ONLY_i + \beta_{27} AGEPCLT6\_6\_12AND13\_17_i +$$
$$\beta_{28} MHONLY_i + \beta_{29} FHONLY_i + \beta_{30} INCOME_i$$

(1)

where $i = 1,......, n$ is the number of households. $Y$ corresponds to the decision to buy drinkable yogurt. Variables are defined in Table 1.

A common characteristic in micro level data (data gathered at consumer level such as at the individual or household level) is a situation where some consumers do not purchase some items during the sampling period and presence of them in the sample creates a zero consumption level for that data period. The data used in this study are gathered at household level and due to that it suffers from zero consumption data. As such we face a censored sample of data. Application of ordinary least squares (OLS) to estimate a regression with a limited dependent variable (such as in a censored sample like ours) usually give rise to biased estimates, even asymptotically (Kennedy, 2003). Removing all observations pertaining to zero purchases and estimating regression functions only for non-zero purchases too creates a bias in the estimates. This phenomenon also is known as *sample selection bias*. Heckman (1979) stated that not adjusting for sample selection may result in biased estimates of the demand parameters. Furthermore, Heckman (1979), discussed the sample selection bias as a specification error, and developed a simple consistent estimation method that eliminates the

specification error for the case of censored samples. It is known as Heckman-type correction procedure.

The first stage of the Heckman-two-step sample selection procedure, involves in decision to purchase drinkable yogurt. It is modeled through a probit model. A binary dependent variable is observed (purchase or not purchase), where purchase is represented by one (1) and not purchase is given by a zero (0). The latent selection equation can be written as follows;

$$Z_h = w'_h \gamma + \varepsilon_h \qquad (2)$$

where $Z_k$ represents a latent selection variable (buy or not to buy type dichotomous variable),

$$Z_h = \begin{cases} 1 \ if \ Z_h > 0 \\ 0 \ if \ Z_h < or = 0 \end{cases} \qquad (3),$$

$w_h$ is a vector of explanatory variables in the latent decision making variable, $\gamma_h$ is a vector of parameters to be estimated in the decision making equation, $\varepsilon_h$ is the error term, and $h = 1,2,.....,N$ is the number of observations (in our work the number of households in the sample) in the sample. Modeling above equation 2 through probit model gives us following relationships;

$$\Pr[Z_h = 1] = \phi(w_h, \gamma) \qquad (4) \text{ and}$$

$$\Pr[Z_h = 0] = 1 - \phi(w_h, \gamma) \qquad (5)$$

where $\phi$ is the normal cumulative probability distribution function (cdf). The first stage estimation provides estimates of $\gamma$ and the inverse of the Mills Ratio (IMR hereinafter). We also generate the associated probability density function (pfd). Inverse of Mills Ratio is calculated taking the ratio of pdf to cdf. Mathematically, it is as follows;

6

$$\text{for } Z_k = 1, \ IMR_h = \frac{\varphi(w_h\hat{\gamma})}{\phi(w_h\hat{\gamma})} \qquad (6),$$

where $\varphi$ represents the probability density function. Inverse mills ratio is a monotone

decreasing function of the probability that an observation is selected into the sample,

$\phi(w_k\hat{\gamma}_k)$ (Heckman, 1979). In particular,

$$\lim_{\phi(Z_{hi})\to 1} IMR_h = 0 \qquad (7)$$

$$\lim_{\phi(Z_{hi})\to 0} IMR_h = \infty \qquad (8)$$

$$\frac{\partial IMR_h}{\partial \phi(Z_h)} < 0 \qquad (9)$$

The calculated IMR, will be used as an additional explanatory variable in the second

stage volume equation, which takes care of the sample selection bias in the data. Second

stage equation is given as follows;

$$E[Y_h \mid Z_h = 1] = X'_h\beta + \alpha\frac{\varphi(w_h\hat{\gamma})}{\phi(w_h\hat{\gamma})} \qquad (10)$$

$$E[Y_h \mid Z_h = 1] = X'_h\beta + \alpha I\hat{MR}_h \qquad (11)$$

where $X_k$ is a vector of explanatory variables considered in the second stage. Importantly,

only observations associated with non-zero observations on $Y_k$ are considered here. The IMR

calculated using information retrieved from first stage probit model is used as an explanatory

variable in the second stage (see equations 10 and 11 above). Presence of a sample selection

bias in data will be communicated through statistical significance of the coefficient

associated with IMR, i.e. $\alpha_k$. If $\alpha_k$ is statistically not different from zero, we conclude that

there is no sample selection bias in the data and result in the following regression model;

$$E[Y_h \mid Z_h = 1] = X'_h\beta_i \qquad (12)$$

It is important to know that the explanatory variables in first stage and second stage equations may or may not be the same. In our work, the price variables in both equations do not. However, rest of the demographic variables is exactly the same in the first stage and second stage.

Choice of explanatory variables in the first and second stage has an implication on the derivation and interpretation of marginal effects associated with variables in the second stage. This is because in the second stage, we have the IMR term augmenting the regular regression function with other explanatory variables. Therefore, in calculating marginal effects, the influence of IMR and its associated regression coefficient on other regression coefficients have to be taken into consideration.

Suppose $X_{kj}$ denote the $j$th regressor that is common to both first stage regressors, $w_k$ and, second stage regressors, $X_j$. Differentiating equation 11 with respect to $j$th regressor, the marginal effect is given by the following relationship (following explanation is borrowed from Saha, Capps and Byrne (1997));

$$\frac{\partial E[Y_h \mid Z_h = 1]}{\partial X_{kj}} = \beta_{ij} + \alpha_i \frac{\partial(\hat{IMR}_{hi})}{\partial X_{hj}} \qquad (13)$$

It is evident from 13 that marginal effect of the $j$th regressor on $Y_{ki}$ consists of two parts: a change in $X_j$ which affects the probability of consuming the commodity (this effect is

represented by $\frac{\partial(\hat{IMR}_{hi})}{\partial X_{hj}}$ in 13); a change in $X_j$ which affects the level of consumption (or

expenditure of consumption) which is conditional upon the household choosing to consume the $i$th commodity (this is represented by $\beta_{ij}$ in 13). The former of the above two expression is important, because the sign and magnitude of the marginal effect depends not only on the $\beta_{ij}$,

but also that of the $\frac{\partial(I\hat{M}R_{hi})}{\partial X_{hj}}$. According to Saha, Capps and Byrne (1997), after some

simplification we get arrive at the following relationship for the Heckman second stage

marginal effects,

$$M\hat{E}_{kj} = \frac{\partial E[y_k \mid Z = 1]}{\partial X_{kj}} = \beta_j - \alpha\gamma_{ij}\{W\gamma IMR_k + (IMR_k)^2\} \qquad (14)$$

In general the marginal effect $M\hat{E}_{kj} \neq \hat{\beta}_j$; however the only case where $M\hat{E}_{kj} = \hat{\beta}_j$ is where

$\hat{\alpha} = 0$ which is a situation where the errors in the first-stage and second-stage estimation

equations have zero covariance. It must be noted that the $M\hat{E}_{kj}$ estimation depends on a local

set of co-ordinates. Therefore, we estimate the $M\hat{E}_{kj}$ at the sample means. Following equation

14 shows this result. For simplicity, let us denote $IMR$ in the letter $\lambda$.

$$M\hat{E}_{kj} \mid_{samplemean} = \hat{\beta}_j - \hat{\alpha}_i\hat{\gamma}_j\{(\overline{W}\hat{\gamma})\overline{\hat{\lambda}} + \overline{\hat{\lambda}}^2\} \qquad (15)$$

where $\overline{W}$ denotes the vector of regressor sample means in the probit equation (the first stage

equation of the Heckman two-step model and

$$\overline{\hat{\lambda}} = \frac{\varphi(\overline{W}\hat{\gamma})}{\phi(\overline{W}\hat{\gamma})} \qquad (16)$$

is the inverse Mills ratio evaluated at those means.

The Heckman two-step demand model for drinkable yogurt can be written as follows:

$$
\begin{aligned}
q_i = {} & \beta_1 + \beta_2 P_i + \beta_3 AGEHH2529_i + \beta_4 AGEHH3034_i + \\
& \beta_5 AGEHH3544_i + \beta_6 AGEHH4554_i + \beta_7 AGEHH5564_i + \beta_8 AGEHHGT64_i + \\
& \beta_9 EMPHHPT_i + \beta_{10} EMPHHFT_i + \beta_{11} EDUHHHS_i + \beta_{12} EDUHHU_i + \\
& \beta_{13} EDUHHPC_i + \beta_{14} REG\_CENTRAL_i + \beta_{15} REG\_SOUTH_i + \\
& \beta_{16} REG\_WEST_i + \beta_{17} RACE\_BLACK_i + \beta_{18} RACE\_ORIENTAL_i + \\
& \beta_{19} RACE\_OTHER_i + \beta_{20} HISP\_YES_i + \beta_{21} AGEPCLT6\_ONLY_i + \\
& \beta_{22} AGEPC6\_12ONLY_i + \beta_{23} AGEPC13\_17ONLY_i + \\
& \beta_{24} AGEPCLT6\_6\_12ONLY_i + \beta_{25} AGEPCLT6\_13\_17ONLY_i + \\
& \beta_{26} AGEPC6\_12AND13\_17ONLY_i + \beta_{27} AGEPCLT6\_6\_12AND13\_17_i + \\
& \beta_{28} MHONLY_i + \beta_{29} FHONLY_i + \beta_{30} INCOME_i + \alpha_i IMR + \varepsilon_i
\end{aligned}
\tag{17}
$$

where $i = 1,......,n$ is the number of observations (households in our work) in the model. $q_i$

corresponds to the quantity of purchase of drinkable yogurt and $P_i$ variable represent the

price of drinkable yogurt. We have defined the variables in the above equation 17 in Table 1.

In the equation 17, *IMR* stands for the inverse Mills ratio and $\alpha_i$ corresponds to the coefficient

associated with *IMR*. Presence of sample selection bias is determined looking at the

significance of $\alpha_i$. If we have sample selection bias, we have to do an adjustment to the

coefficient estimates in the second stage estimation in trying to get at correct marginal

effects. Procedure to adjust for marginal effects was elaborated in the preceding section.

As such, we will calculate marginal effects associated with each explanatory variable.

The level of significance we will be using in this study is 0.05. We further conduct an *F*-test

for demographic variable categories to find statistically significant demographics.

**Results and Discussion**

Preliminary investigations using transactions data reveal that the market penetration

for drinkable yogurt is 19 percent. The average at-home quantity of DY consumed is 133

ounces per household per year and the average price is $0.14 per ounce. Factors affecting the

probability of purchase of drinkable yogurt are price of drinkable yogurt, income of the

household head, employment status of the household head, education status of the household head, region, race, age and presence of children, and gender of the household head. Factors affecting the volume of purchase of drinkable yogurt are price of drinkable yogurt, region, race, and gender of the household head. The own-price elasticity of demand for drinkable yogurt is estimated to be -0.68. A significant sample selection bias is observed in considering the demand for drinkable yogurts.

**References:**

Beverage Marketing Corporation, 2008

Convenient Store News, (2010). "Yogurt Drinks Poised for Further Growth". Online accessed www.csnews.com

Davis, Christopher G., D. Dong, D.P. Blayney, and A. Owens. (2010) "An Analysis of U.S. Household Dairy Demand." TB-1928, U.S. Department of Agriculture, Economic Research Service

Fuller, W.A., and G.E. Battese. (1974). "Estimation of Linear Models with Crosses-Error Structure". *Journal of Econometrics* 2: 67-78

Heckman, J. J., (1979). "Sample Selection Bias as a Specification Error". *Econometrica* 47: 153-161

Ozer, B. H., and H. A. Kirmaci. (2010). "Functional Milks and Dairy Beverages". *International Journal of Dairy Technology* 63(1):1-15

Thompson, J.L., K. Lopetcharat, and M.A.Darke. (2007). "Preferences for Commercial Strawberry Drinkable Yogurts Among African American, Caucasian, and Hspanic Consumers in the United States". *Journal of Dairy Science* 90(11): 4974-4987

Saha, A., O.Capps. Jr, et al. (1997). "Calculating Marginal Effects in Dichotomous-Continuous Models." Applied Economic Letters **4**(3): 181-185.

Kennedy, P. (2003). Limited Dependent Variables. A Guide to Econometrics, MIT Press.

**Table 1 Description of the Right-Hand Side Variables Used in the Econometric Analysis**

| Variable | Explanation |
|---|---|
| PRICE | Price of Drinkable yogurt |
| *AGEHHLT25* | *Age of Household Head less than 25 years (Base category)* |
| AGEHH2529 | Age of Household Head between 25-29 years |
| AGEHH3034 | Age of household Head between 30-34 years |
| AGEHH3544 | Age of household Head between 35-44 years |
| AGEHH4554 | Age of household Head between 45-54 years |
| AGEHH5564 | Age of household Head between 55-64 years |
| AGEHHGT64 | Age of household Head greater than 64 years |
| *EMPHHNFP* | *Household Head not employed for full pay (Base category)* |
| EMPHHPT | Household Head Part-time Employed |
| EMPHHFT | household Head Full-time Employed |
| *EDUHHLTHS* | *Education of Household Head: Less than high school (Base category)* |
| EDUHHHS | Education of Household Head: High school only |
| EDUHHU | Education of Household Head: Undergraduate only |
| EDUHHPC | Education of Household Head: Some post-college |
| *EAST* | *Region: East (Base category)* |
| MIDWEST | Region: Central (Midwest) |
| SOUTH | Region South |
| WEST | Region West |
| *WHITE* | *Race White (Base category)* |
| BLACK | Race Black |
| ASIAN | Race Oriental |
| RACE_OTHER | Race Other (non-Black, non-White, non-Oriental) |
| *HISP_NO* | *Non-Hispanic Ethnicity (Base category)* |
| HISP_YES | Hispanic Ethnicity |

**Table 1 Continued….**

| Variable | Explanation |
|---|---|
| *NPCLT_18* | *No Child less than 18 years (Base category)* |
| AGEPCLT6_ONLY | Age and Presence of Children less than 6-years |
| AGEPC6_12ONLY | Age and Presence of Children between 6-12 years |
| AGEPC13_17ONLY | Age and Presence of Children between 13-17 years |
| AGEPCLT6_6_12ONLY | Age and Presence of Children less than 6 and 6-12 years |
| AGEPCLT6_13_17ONLY | Age and Presence of Children less than 6 and 13-17 years |
| AGEPC6_12AND13_17ONLY | Age and Presence of Children between 6-12 and 13-17 years |
| AGEPCLT6_6_12AND13_17 | Age and Presence of Children less than 6, 6-12 and 13-17 years |
| *FHMH* | *Household Head both Male and Female (Base category)* |
| MHONLY | Household Head Male only |
| FHONLY | Household Head Female only |