THREE ESSAYS ON (MODELING) HOUSEHOLD FOOD PURCHASE BEHAVIORS

by

SHENGFEI FU

(Under the Direction of Wojciech J. Florkowski)

ABSTRACT

This dissertation consists of three essays investigating household food purchase behaviors, focusing on modeling household binary purchase choices and expenditure decisions. The findings reveal factors that are influential on the formation of healthy and/or unhealthy dietary choices and provide insights for producers, retailers, and public health policymakers.

The first essay proposes a new estimator for multivariate binary response data, a data feature of growing interest in the study of consumer behavior. This study considers binary responses as being generated from a truncated multivariate discrete distribution. The new estimator is shown to have attractive properties through Monto Carlo simulations and empirical applications. Comparisons are made to the traditional multivariate probit model. Because multivariate binary response modeling is frequently required in areas such as marketing, household behavior, crop selection, and conservation practices, among others, findings are of interest to both econometricians and practitioners.

The second essay investigates the effects of demographic and socio-economic factors as well as outmigration, a special issue in Poland, on the consumption of tobacco and alcohol. This study takes advantage of second-hand survey data collected from a household panel by Poland's Main Statistical Office (GUS) that is not publicly available. Due to the addictive nature of tobacco

and alcohol, this study uses a censored system to model the correlated consumption of tobacco and alcohol. Findings provide insights for the reduction and prevention of tobacco and alcohol use.

The third essay provides a holistic profile of fresh produce choices and expenditures, including expenditure on fresh produce, frequency of purchase, variety of selection, and use of deals and coupons. A profile of consumers by consumer group was developed using 2014 Nielsen Homescan panel. This study intends to present a holistic picture of consumer disadvantage in terms of fresh produce consumption and take an all-inclusive approach so as to seek out commodities as well differences in fresh produce shopping behaviors across four consumer groups.

INDEX WORDS: consumer behavior, food choice, multivariate binary responses, discrete normal distribution, multivariate probit, smoke, drink, tobacco and alcohol consumption, censored system, Heckman's sample selection model, multivariate sample selection model, worker migration, fresh produce, fruit and vegetable consumption, at-home consumption, disadvantaged consumers

THREE ESSAYS ON (MODELING) HOUSEHOLD FOOD PURCHASE BEHAVIORS

by

SHENGFEI FU

B. A., University of International Relations, China, 2009

M. S., The University of Georgia, USA, 2012

M. S., The University of Georgia, USA, 2014

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

THREE ESSAYS ON (MODELING) HOUSEHOLD FOOD PURCHASE BEHAVIORS

by

SHENGFEI FU

Major Professor: Wojciech J. Florkowski

Committee: J. Scott Shonkwiler

Gregory J. Colson

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2016

DEDICATIONS

Dedicated to my beloved and supporting parents and parents-in-law, my lovely husband, my adorable little son, and my dear brother.

For their endless love, support, encouragement and sacrifices

ACKNOWLEDGEMENTS

As I take the final step to finish the arduous journey of pursuing Ph.D., I would like to take this opportunity to express my genuine gratitude to all those who have made it real. This list is no way exhaustive or exclusive.

I am privileged and honored to have Dr. Florkowski as my major advisor. Dr. Florkowski has continuously supported and guided me throughout my program. His wisdom, patience, encouragement, support and professional knowledge are greatly appreciated. Dr. Florkowski not only guides me in my research and dissertation, but also generously supports me to build my career. To him, I am eternally grateful, and I hope that the relationship we have built can continue in the future.

My sincere gratitude goes to Dr. Shonkwiler, my committee and co-author of my first essay. His knowledge, by no means limited to, Econometrics and modeling skills in Matlab is immense and his enthusiasm contagious. Dr. Shonkwiler's intelligent insights, unlimited advice and support help me finish my dissertation and begin a new chapter of my life.

I am deeply grateful to Dr. Colson for serving on my committee and giving me advice and support. As the major advisor of my master program, Dr. Colson mentored me thoroughly through the entire process. His professional knowledge and generous help continue to benefit me beyond my master program.

My sincere gratitude also goes to the Department of Agricultural and Applied Economics for financial support, to the professors from whom I learned knowledge and received generous help, to the staff members who made my study experience smoother, and to fellow graduate

students for their support and friendship. I will not mention their names one by one because it will take more than one page. But I would like to let them know that I will never forget every moment that we ever shared. Their professional knowledge, friendliness, and support enrich my journal.

Finally, my deepest thanks and love goes to all of my family. None of my accomplishments is possible without their endless support, love, encouragement, and sacrifices. Words cannot convey how much each of them means to me. I am privileged to have them in my life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

An escalating global epidemic of overweight and obesity paradoxically coexist with undernutrition (WHO 2016a). Institutions and governments are promoting healthy diet as a main venue for addressing this issue. Understanding how socio-economic, demographic and behavioral factors influence food choices is key to developing a solution to effectively promote healthy diet choices. This dissertation attempts to add efforts to this agenda from different yet related perspectives. This dissertation consists of three essays investigating household food purchase behaviors, focusing on modeling household binary purchase choices and expenditure decisions.

1.1 Modeling Multivariate Binary Responses

Consumers make vast discrete choices, such as where to shop, whether to purchase foods, or which brand to purchase. The model for binary choice is the fundamental pillar of analysis of discrete choice (Greene 2009). There is growing interest concerning the analysis of correlated binary data in the study of consumer behavior. While the multivariate probit model is widely regarded as the preferred estimator of correlated binary response variables, its use is hurdled by the computational burden. Exact maximum likelihood estimation (MLE) of the multivariate probit model involves evaluation of M-variate normal cumulative density function (cdf), which cannot be performed analytically or with quadrature. Thus considerable research has been undertaken to evaluate simulation estimators for this model. It appears that the Geweke-Hajivassiliou-Keane (GHK) smooth recursive simulator (Geweke 1989; Hajivassiliou and McFaddem 1998; and Keane 1994)

dominates all the simulation methods proposed to date. Yet this estimator is computationally demanding in large systems and results may be sensitive to the number of random draws. The first essay of this dissertation proposes an alternative estimator based on discrete normal distribution (Kemp 1997). The new estimator is derived and its attractive properties are mathematically proven and empirically illustrated in Chapter 2. Because multivariate binary response modeling is frequently required in areas such as marketing, household behavior, crop selection, and conservation practices, among others, it is the author's believe that the findings are of interest to both econometricians and practitioners.

1.2 Household Consumption of Tobacco and Alcohol in Poland

Tobacco and alcohol are not food items, but their expenditure affect the allocation of food dollars. Additionally, the consumption of tobacco and the excessive use of alcohol products adversely impact health outcome. Motivated by the increasing interest in promoting healthy lifestyle, it is in of utmost concern to understand the determinants of tobacco and alcohol consumption.

Tobacco and alcohol-related harm are two high-priority public health challenges facing Europe. Percent of consumer expenditures spent on alcoholic beverages and tobacco that were consumed at home in Poland were 7.6% in 2014 (ERS, USDA, 2015), ranking 8th among 86 selected countries being reported. With the growing attention to tobacco- and alcohol-related social and health problems and public policy campaigns against tobacco and alcohol use, an analysis of the determinants of household tobacco and alcohol consumption remains important.

Researchers have long hypothesized a two-stage choice process where consumers first decide whether to buy a commodity, and then choose the amount to purchase (e.g. Bettman 1979; Gensch 1987; Shocker et al. 1991; Wright and Barbour 1977). The analysis of tobacco and alcohol

consumption in Poland follows that well-accepted hypothesis. Additionally, the addictive nature as well as observed culture of smoking and drinking suggests that it is more appropriate to model the consumption of tobacco and alcohol as a system (Pierani and Tiezzi 2009). A multivariate sample selection model is extended from the traditional Heckman's sample selection model (Heckman 1976, 1979) to accommodate empirical analysis. The current analysis reveals household features that are associated with higher consumption of tobacco and alcohol in terms of purchase probabilities and level of expenditures. Findings provide insights for the reduction and prevention of tobacco and alcohol use.

1.3 A Holistic Picture of At-home Consumption of Fresh Produce

Nutrition disorders, either insufficient intake or overconsumption of certain food or nutrition are related to adverse health outcomes such as noncommunicable diseases (NCDs), including diabetes, heart disease, stroke and cancer. Nutrition disorder can be particularly serious in children since they interfere with growth and development, and may predispose to many health problems, such as infection and chronic disease. Healthy diet helps protect against malnutrition in all its forms, as well as NCDs. But the increases in production of processed food, rapid urbanization and changing lifestyles have led to a shift in dietary patterns. People are now consuming more foods high in energy, fats, free sugars or sodium, and many do not eat enough fruit, vegetables and dietary fiber (WHO 2016b).

Disadvantaged groups are least likely to have fruit and vegetable intakes that are consistent with healthy eating messages (Giskes et al. 2002; Mishra et al., 2002; and Turrell et al., 2002). Many studies have set out to investigate food shopping patterns of one specific group of disadvantaged consumers, for example, the low-income (e.g. Clifton 2004), the elderly (e.g.

Wilson, Alexander and Lumbers 2004), women (e.g. Herman et al., 2008), and ethnic minorities (e.g. Zenk 2005). However, studies that take a more holistic view have been lacking.

The third essay of this dissertation intends to present a holistic picture of consumer disadvantage and take an all-inclusive approach so as to seek out commonalities as well as differences in fresh produce shopping behavior across different characteristics of disadvantaged consumers. Such findings provide key insights to developing a solution to the rising rate of obesity in the United States. Socio-economic and demographic features indicating consumer disadvantage were given special attention in the third essay.

The remainder of this dissertation is organized as follows: Chapters 2 to 5 present each of the three essays and Chapter 5 concludes with summary and discussion.

CHAPTER 2

A NEW ESTIMATOR FOR MULTIVARIATE BINARY DATA

There is growing interest concerning the analysis of correlated binary data in the study of consumer behavior. The multivariate probit model (MVP) is widely regarded as the preferred estimator of correlated binary response variables. Unfortunately, exact maximum likelihood estimation of the multivariate probit requires the evaluation of an M^{th} order integral when there are M correlated binary responses. Simulation estimators are computationally demanding and results may be sensitive to the number of random draws. This study proposes a new estimator for multivariate binary response data and considers binary responses as being generated from a truncated multivariate discrete distribution. Specifically, the discrete normal probability mass function, which has support on all integers, is extended to a multivariate form. Truncating this point probability mass function below zero and above one results the multivariate binary discrete normal distribution. This distribution has a number of attractive properties. Monte Carlo simulation and empirical applications are performed to show the properties of this new estimator; comparisons are made to the traditional multivariate probit model. Because multivariate binary response modeling is frequently required in areas such as marketing, household behavior, crop selection, and conservation practices, among others, the findings are of interest to both econometricians and practitioners.

2.1 Introduction

There is growing interest concerning the analysis of correlated binary data in the study of consumer behavior. The purchase of specific products and/or brands, the choice of shopping venues, and the selection of certain activities are amenable to binary response modeling.

The multivariate probit model is widely regarded as the preferred estimator of correlated binary response variables. Unfortunately, exact maximum likelihood estimation of the MVP requires the evaluation of an M^{th} order integral when there are M correlated binary responses. Thus considerable research has been undertaken to evaluate simulation estimators for this model. It appears that the Geweke-Hajivassiliou-Keane (GHK) smooth recursive simulator (Geweke 1989; Hajivassiliou and McFaddem 1998; and Keane 1994) dominates all the simulation methods proposed to date. Yet this estimator is computationally demanding in large systems and results may be sensitive to the number of random draws.

This study considers binary responses as being generated from a truncated multivariate discrete normal distribution. The discrete normal distribution – as defined by Kemp (1997) – has support on all integers. The discrete normal probability mass function is extended to a multivariate form. Doubly truncating this joint probability mass function below zero and above one results the multivariate binary discrete normal distribution for a system of binary response variables. Maximum likelihood estimation is straightforward because for M response variables, only 2^M support points need to be evaluated to obtain the normalizing factor for the multivariate binary normal probability mass function. This new estimator for multivariate binary response data is termed as Multivariate Binary Discrete Normal (MBDN) Estimator.

The multivariate binary discrete normal distribution has a number of attractive properties: it is a member of the quadratic exponential family; analogous to continuous normal distribution,

its marginal and conditional distributions are also discrete normal distribution; it nests a system of independent binary logits; it does not require conditioning to eliminate nuisance parameters; and exact maximum likelihood estimation is feasible since the normalizing factor only requires the evaluation of 2^M support points for a system of M response variables. Monte Carlo simulation and empirical application are performed to show the properties of this new estimator and comparisons are made to the traditional MVP model.

The remainder of this chapter is organized as follows: Section 2.2 derives the new estimator from the multivariate binary discrete normal distribution and discusses its maximum likelihood estimation. Section 2.3 explains the simulation setup and compares MBDN estimates to the traditional MVP model. Section 2.4 reports empirical applications. Finally, section 2.5 concludes with a discussion.

2.2 Multivariate Binary Discrete Normal Estimator

2.2.1 The Discrete Normal Distribution

Kemp (1997, p.224) characterizes the probability mass function (pmf) of a discrete normal random variable Y with parameters (λ, q) as

(2.1) $$P(Y = y) = \frac{\lambda^y q^{y(y-1)/2}}{\sum_{x=-\infty}^{\infty} \lambda^y q^{y(y-1)/2}}$$, y =…, −2, −1, 0, 1, 2, …; λ>0 and 0< q <1.

The discrete normal distribution has a number of attractive properties: 1) the discrete normal distribution is analogous to the normal distribution in that it is the only two-parameter discrete distribution on (−∞, ∞) for which the first two moment equations are the maximum-likelihood equations; 2) the distribution is unimodal like the normal distribution; and 3) the distribution is log-concave like the normal distribution (Kemp 1997, p.225).

Let $\lambda = \exp((\mu - 0.5)/\sigma^2)$ and $q = \exp(-1/\sigma^2)$ such that $-\infty < \mu < \infty$ and $\sigma^2 > 0$. The discrete normal may now be represented as

(2.2) $$P(Y = y) = \frac{\exp(-0.5(y-\mu)^2/\sigma^2)}{\sum_Y \exp(-0.5(Y-\mu)^2/\sigma^2)},$$

where y = … , −2, −1, 0, 1, 2, …; $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

The discrete normal can handle binary data by doubly truncating outcomes below zero and above one. Since location and not scale is of interest, the variance is set $\sigma^2=1$, then P(Y=0)=1−P(Y=1) and

(2.3) $$P(Y = 1) = \frac{\exp(-0.5(1-\mu)^2\ )}{\exp(-0.5(0-\mu)^2)+\exp(-0.5(1-\mu)^2)} = \frac{\exp(\mu-0.5)}{1+\exp(\mu-0.5)}.$$

It is obvious that this model is indistinguishable from the conventional binary logit model under the parameterizations μ=Xβ when X contains a constant. This feature indicates that the binary discrete normal distribution may be more applicable to data with thicker tails than the normal distribution.

The univariate discrete normal distribution can be generalized to the multivariate case of M integer responses using the following representation for a single observation:

(2.4) $$\mathrm{P}\ (Y_1 = y_1, Y_2 = y_2, \ldots, Y_M = y_M) = \frac{\exp(-0.5(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}))}{\sum_{Y_1,\ldots,Y_M} \exp(-0.5(\boldsymbol{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}))}$$

y_m = …, −2, −1, 0, 1, 2, … ∀ m=1, 2, … M.

where $\boldsymbol{\Sigma}$ is assumed to be a positive definite M x M symmetric matrix and $\mathbf{y}$ and $\boldsymbol{\mu}$ are M x1 vectors. The summation term in the denominator represents all points of support of the distribution.

Discrete Normal belongs to the Exponential Family

To show that the discrete normal distribution is a member of the exponential family recall the pmf in terms of parameters μ and σ^2 as expressed in Equation (2). It can be rewritten as

(2.5) $$P(Y=y)=\frac{\exp(-\mu^2/2\sigma^2)\exp(y\mu/\sigma^2-y^2/2\sigma^2)}{\exp(-\mu^2/2\sigma^2)\sum_Y \exp(Y\mu/\sigma^2-Y^2/2\sigma^2)}.$$

Then the canonical representation of the pmf is

(2.6) $$P(Y=y)=\frac{\exp(\theta_1 y+\theta_2 y^2)}{\sum_Y \exp(\theta_1 Y+\theta_2 Y^2)}$$, where $\theta_1=\mu/\sigma^2$ and $\theta_2=-1/2\sigma^2$.

This representation has the form a(**θ**)exp(**θ**´T(**y**)) where a(**θ**) is the normalizing factor and T(**y**) = [y y^2]', then the discrete normal distribution is a member of the exponential family. Consequently if defining κ(**t**) = ln(exp(tY)a(**θ**)$^{-1}$), then the derivatives of κ(**t**) with respect to t evaluated at t=0 are the cumulants of y. This is particularly important because a truncated distribution from the exponential family merely has the domain of Y restricted to a subspace and remains a member of this family (Lindsey 1996, p.37).

By analogy to the continuous multivariate normal distribution, the joint discrete normal pmf has, upon canceling the common term exp(–1/2μΣ^{-1}μ'), the canonical form as below:

(2.7) $$P(Y=y)=\exp\left[\sum_{i=1}^{k} c_i(\Theta)T_i(y)-\ln a(\Theta)\right],$$

where Θ is a k=m+m(m+1)/2 parameter vector ; c(Θ)´={μ´Σ^{-1},-1/2 Σ^{11}, -Σ^{1m}, …, -1/2Σ^{22}, -Σ^{23} ,…, -1/2Σ^{mm} } where Σ^{ij}=[Σ^{-1}]$_{ij}$; T(**y**)´=={**y**', y_1^2, y_1y_2, …, y_1y_m, y_2^2, y_{23}, …, y_m^2}; and $a(\Theta)=\sum_{Y_m}\ldots\sum_{y_1}\exp(c(\Theta)'T(Y))$ is a finite, real-valued function which does not depend on y. Therefore, this joint pmf is a member of the exponential family.

The MBDN model is a special case of the quadratic exponential models developed by Prentice and Zhao (1990) and Fitzmaurice and Laird (1993) for the analysis of multivariate binary observations. As such, the maximum likelihood estimator can possess attractive properties even

under distributional misspecification (Gourieroux, Monfort, and Trognon, 1984). Worthy of note is that as a result of double truncation, the variance of the MBDN model is not identified because y_m^2 in T(**y**) reduces to be y_m, given the only possible values for y_m are 0 and 1. This is why the diagonal elements of **Σ** in Equation (2.4) and later relevant equations are constrained to unity for identification.

2.2.2 Multivariate Binary Discrete Normal Distribution

For multivariate binary responses, the random variables are doubly truncated below zero and above one so that the support becomes $y_m = 0, 1 \; \forall \; m=1, 2, \dots , M$. The diagonal elements of **Σ** are constrained to unity for identification, and under independence ($\sigma_{ij}=0; i \neq j$), it clearly nests independent binary logit models. This model is termed as the multivariate binary discrete normal (MBDN) estimator and its probability mass function is

$$(2.8) \quad P(Y_1 = y_1, Y_2 = y_2, \dots, Y_M = y_M) = \frac{\exp(-0.5(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}))}{\sum_{Y_1,\dots,Y_M} \exp(-0.5(\boldsymbol{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}))}$$

$y_m = 0, 1 \; \forall \; m=1, 2, \dots M.$

Note that because of the truncation, the number of points of support needed to calculate the normalizing factor in the denominator amounts to just 2^M. Thus, exact maximum likelihood estimation of multivariate binary response models is quite feasible on systems with M as large as 20.

2.2.3 Maximum Likelihood Estimation of MBDN Estimator

It can be shown that for the multivariate binary discrete normal distribution the first two moment equations are the maximum likelihood equations when regressors are identical across equations. The multivariate binary discrete normal joint probability mass function can be rewritten as

(2.9) $$P(\boldsymbol{Y}=\boldsymbol{y})=\frac{\exp\left(0.5\boldsymbol{y}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+0.5\boldsymbol{\mu}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{y}-0.5\boldsymbol{y}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right)}{\sum_{\boldsymbol{Y}}\exp\left(0.5\boldsymbol{Y}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+0.5\boldsymbol{\mu}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}.$$

The log likelihood for the i[th] observation (suppressing subscripts) is as follows:

(2.10) $$l_{(i)}=0.5\boldsymbol{y}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+0.5\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}-0.5\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}-\ln\left(\sum_{\boldsymbol{Y}}\exp\left(0.5\boldsymbol{Y}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+0.5\boldsymbol{\mu}'\,\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)\right).$$

Let's first illustrate through a simple case where **μ** is a column vector of constants. Solving for the maximum likelihood estimators obtains:

(2.11) $$\frac{\partial l_{(i)}}{\partial\boldsymbol{\mu}}=\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}-\frac{\sum_{\boldsymbol{Y}}\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}{\sum_{\boldsymbol{Y}}\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)},\text{ and}$$

(2.12) $$\sum_i\frac{\partial l_{(i)}}{\partial\boldsymbol{\mu}}=\sum_i\boldsymbol{y}_i\,\boldsymbol{\Sigma}^{-1}-\frac{n\sum_{\boldsymbol{Y}}\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}{\sum_{\boldsymbol{Y}}\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}=0.$$

The first-order condition for the estimator of **μ** is then:

(2.13) $$\bar{\boldsymbol{y}}=\frac{\sum_{\boldsymbol{Y}}\boldsymbol{Y}\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}{\sum_{\boldsymbol{Y}}\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}$$

which is the definition of the mean of the vector **y** over the sample. Note that **μ** is not the estimator of $\bar{\boldsymbol{y}}$ even when **Σ** is an identity matrix. This result stems from the representation in Equation (2.3).

Solving for the ML estimator of **Σ** yields:

(2.14) $$\frac{\partial l_{(i)}}{\partial\boldsymbol{\Sigma}}=-0.5\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}\boldsymbol{y}'+\boldsymbol{y}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}+0.5\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}$$

$$-\frac{\sum_{\boldsymbol{Y}}\left(-0.5\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}\boldsymbol{Y}'+\boldsymbol{Y}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}+0.5\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\right)\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}{\sum_{\boldsymbol{Y}}\exp\left(\boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'-0.5\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'\right)},\text{ and}$$

(2.15) $$\sum_i\frac{\partial l_{(i)}}{\partial\boldsymbol{\Sigma}}=0.5\boldsymbol{\Sigma}^{-1}\left[\left(\sum_i\boldsymbol{y}_i\boldsymbol{y}_i'-n\frac{\sum_{\boldsymbol{Y}}\boldsymbol{Y}\boldsymbol{Y}'\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}{\sum_{\boldsymbol{Y}}\exp\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\right)}\right)\right.$$

$$-\boldsymbol{\mu}\left(\sum_i \boldsymbol{y}_i' - n\frac{\sum_{\boldsymbol{Y}} \boldsymbol{Y}\exp(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})}{\sum_{\boldsymbol{Y}}\exp(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})}\right) - \left(\sum_i \boldsymbol{y}_i' - n\frac{\sum_{\boldsymbol{Y}} \boldsymbol{Y}\exp(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})}{\sum_{\boldsymbol{Y}}\exp(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})}\right)' \boldsymbol{\mu}\Big]\boldsymbol{\Sigma}^{-1} = 0.$$

Upon simplification and using the previous first order condition in Eq. (2.9) and defining $\boldsymbol{S}_{yy} = \sum_i \boldsymbol{y}_i\boldsymbol{y}_i'/n$, it follows that the ML estimator of $\boldsymbol{\Sigma}$ satisfies

(2.16) $$\boldsymbol{S} = \frac{\sum_{\boldsymbol{Y}} \boldsymbol{Y}\boldsymbol{Y}'\exp(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})}{\sum_{\boldsymbol{Y}}\exp(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})} \equiv \boldsymbol{S}_{yy},$$

which is the definition of the un-centered sample variance-covariance matrix $\mathbf{S_{yy}}$. With Equations (2.13) and (2.16), it is obvious that the MLE estimate of the correlation matrix from MBDN model equals sample correlation matrix, since $\boldsymbol{S} - \bar{y}\bar{y}' = E(yy') - Ey(Ey)'$.

Let's generalize by defining $\boldsymbol{\mu_i} = \mathbf{X_i}\boldsymbol{\beta}$, $\mathbf{X_i}$=diagonal($\mathbf{x_{1i}}$, $\mathbf{x_{2i}}$, …, $\mathbf{x_{Mi}}$), and $\boldsymbol{\beta}$=[$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, …, $\boldsymbol{\beta}_M$]′, where $\mathbf{x_{mi}}$ ∀m=1, 2, …, M is a 1 x k_m vector, containing a constant term and (k_m −1) explanatory variables for the mth response. $\mathbf{X_i}$ is a (M x $\sum_{m=1}^{M} k_m$) matrix and $\boldsymbol{\beta}$ is a $\sum_{m=1}^{M} k_m$ elements column vector. Then the first-order conditions with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are as follows, respectively:

(2.17) $$\frac{1}{n}\sum_i(\boldsymbol{y}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i) = \frac{1}{n}\sum_i E(\boldsymbol{y}_i')\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i,$$ and

(2.18) $$\boldsymbol{S} = \frac{1}{n}\sum_i \frac{\sum_{\boldsymbol{Y}} \boldsymbol{Y}\boldsymbol{Y}'\exp(\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})}{\sum_{\boldsymbol{Y}}\exp(\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})} + \boldsymbol{A} + \boldsymbol{A}' = \boldsymbol{S}_{yy} + \boldsymbol{A} + \boldsymbol{A}',$$

where $\boldsymbol{A} = \frac{1}{n}\sum_i \boldsymbol{\mu}_i\left(\boldsymbol{y}_i' - \boldsymbol{E}(\boldsymbol{y}_i)\right)$ and $E(\boldsymbol{y}_i) = \frac{\sum_{\boldsymbol{Y}} \boldsymbol{Y}\exp(\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})}{\sum_{\boldsymbol{Y}}\exp(\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}-0.5\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})}$.

Equation (2.17) is essentially a collection of $\sum_{m=1}^{M} k_m$ equations. Let $\boldsymbol{x}_{\boldsymbol{mi}}^*$ be a 1 x (k$_m$-1) vector containing the (k_m −1) explanatory variables for the mth response. Then $\mathbf{X_i}$ can be re-organized such that $\widetilde{\boldsymbol{X}_{\boldsymbol{\iota}}} = \begin{bmatrix} \boldsymbol{I}_{\boldsymbol{M}} & \\ & \boldsymbol{X}_{\boldsymbol{i}}^* \end{bmatrix}$ where $\boldsymbol{X}_{\boldsymbol{i}}^* = diag(\boldsymbol{x}_{\boldsymbol{1i}}^*, \ldots, \boldsymbol{x}_{\boldsymbol{Mi}}^*)$. The new-ordered first M

equations then reduce to the first moment condition as in Equation (2.13). The remaining $(\sum_{m=1}^{M} k_m - M)$ equations provide additional information that will be explored later:

(2.19) $\frac{1}{n}\sum_i \left(\boldsymbol{y_i'} - \boldsymbol{E}(\boldsymbol{y_i'})\right)\boldsymbol{\Sigma^{-1}}\boldsymbol{X_i^*}) = \mathbf{0}$.

Note in Equation (2.18) the first part of the right-hand side (RHS) is the un-centered sample variance-covariance. The two additional components, **A** and **A**', form a symmetric M x M matrix, which becomes zero when X is identical across equations. This is easily seen when **μ** is a vector of constants. A proof is given in Appendix to show that this is also the case when X is identical across the M response variables.

2.2.4 Marginal Effects

In many economic studies, one is interested in the marginal effect of an explanatory variable. Using the definition of the mean as given in Equation (13), the derivatives of E(**y**) with respect to the **μ** is derived to be an M x M matrix:

(2.20) $$\frac{\boldsymbol{\partial E(y)}}{\boldsymbol{\partial \mu}} = \begin{bmatrix} \frac{\partial E(y_1)}{\partial \mu_1} & \frac{\partial E(y_1)}{\partial \mu_2} & \cdots & \frac{\partial E(y_1)}{\partial \mu_M} \\ \frac{\partial E(y_2)}{\partial \mu_1} & \frac{\partial E(y_2)}{\partial \mu_2} & \cdots & \frac{\partial E(y_2)}{\partial \mu_M} \\ \vdots & \vdots & & \vdots \\ \frac{\partial E(y_M)}{\partial \mu_1} & \frac{\partial E(y_M)}{\partial \mu_2} & \cdots & \frac{\partial E(y_M)}{\partial \mu_M} \end{bmatrix} = \frac{\sum_{\boldsymbol{Y}} \boldsymbol{Y}\left(\boldsymbol{\Sigma^{-1}}[\boldsymbol{Y}-\boldsymbol{E}(\boldsymbol{y})]\right)' \exp\left((\boldsymbol{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma^{-1}}(\boldsymbol{Y}-\boldsymbol{\mu})\right)}{\sum_{\boldsymbol{Y}} \exp\left((\boldsymbol{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma^{-1}}(\boldsymbol{Y}-\boldsymbol{\mu})\right)}.$$

Let **μ= Xβ** as previously defined and $\boldsymbol{\beta_{mk}}$ be the parameter of the k[th] regressor in **X** for the m[th] response. The derivative of the **μ** with respect to x_k is an M-element column vector:

(2.21) $$\frac{\boldsymbol{\partial \mu}}{\boldsymbol{\partial} x_k} = \left[\frac{\partial E(\mu_1)}{\partial x_k} \quad \frac{\partial E(\mu_2)}{\partial x_k} \quad \ldots \quad \frac{\partial E(\mu_M)}{\partial x_k}\right]' = [\mathbf{1_1}(x_k)\beta_{1k} \quad \mathbf{1_2}(x_k)\beta_{2k} \quad \ldots \quad \mathbf{1_M}(x_k)\beta_{Mk}]',$$

where $\mathbf{1_m}(x_k) = \begin{cases} 1 \; if \; x_k \in x_m \\ 0 \; otherwise \end{cases} \qquad \forall \; m = 1, 2, \ldots, M.$

The marginal effects of x_k on the M response variables are then obtained by multiplying Equations (2.20) and (2.21) according to the chain rule of derivatives.

Note the marginal effects involve all M sets of regressors if there are common variables. The marginal effects are the addition of direct effects on the response variable and indirect effects through the other (M-1) response variables. When event 1 indicates purchase, the marginal effect of x_k is interpreted as change in the probability of purchasing product m, corresponding to a one-unit change in x_k. This study reports average marginal effects.

A straightforward way to derive marginal effects under the MVP model is to use the univariate unconditional distributions. As Mullahy (2011) points out, while such aggregation approaches may be informative for some purposes, it should be emphasized that they fail fundamentally to represent the properties of the underlying probability structure of the multivariate model. More appropriate marginal effects are based on joint conditional probabilities or probabilities conditional on subvectors of **y**. This complicates the computation of marginal effects because there is an ambiguity in the conditional distributions. Given the dimension of M response variables, there are 2^M probability outcomes based on the joint conditional probabilities. And the outcomes based on probabilities conditional on subvectors of **y** mount to $\sum_{k=1}^{M-1}\binom{M}{k}2^k$ (which is 4, 18, 64 for M=2, 3, 4, respectively). See Mullahy (2011) for a general analytical formula for such marginal effects. As a comparison, the marginal effects under MBDN are much easier to compute, since an easy to implement formula of the expected mean functions are given by the definition. This adds another attractive property to the MBDN model.

As previously pointed out, there are eventually 64 outcome combinations based on joint or conditional distributions for the 4-variate normal distribution in question. For simplicity, this study shows a subset of marginal effects for illustration through numerical iteration.

2.3 Monte Carlo Simulations

Monte Carlo simulations are performed to show the properties of the MBDN estimator and provide comparisons to the traditional MVP model. The first set of simulations generates multivariate binary responses assuming that they are drawn from an underlying normal distribution. This study will compare the average marginal and conditional probabilities from the multivariate binary discrete normal to those estimated by the MVP model. Thus the study should be able to provide some guidelines as to the applicability of the multivariate binary discrete normal when the data are known to come from a multivariate normal distribution. Next the study will carry out simulations where the binary responses come from an underlying multivariate t-distribution. The idea here is that the multivariate binary discrete normal may actually outperform the MVP model in such circumstances.

2.3.1 Data Generating Process

Consider the M-equation multivariate binary model, where, for convenience, the individual observation index is omitted:

(2.22) $y_m^* = \boldsymbol{\beta}_m' \boldsymbol{X}_m + \varepsilon_m, m = 1, \dots, M;\ \ y_m = 1\ if\ y_m^* > 0\ and\ 0\ otherwise.$

In the first set of simulations, ε_m, m=1, …, M are error terms distributed as multivariate normal (MVN), each with a mean of zero, and variance-covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ has values of 1 on the leading diagonal and correlations $\sigma_{jk} = \sigma_{kj}$ as off-diagonal elements. A random draw from an MVN distribution can be obtained using the Cholesky decomposition of $\boldsymbol{\Sigma}$, the lower triangular factor **A**, for which $\mathbf{AA}' = \boldsymbol{\Sigma}$, and a vector of univariate normal draws, z. Specifically, Z=**A**z is a random draw from the MVN distribution with mean vector zero and covariance matrix $\boldsymbol{\Sigma}$.

In the second set of simulations, ε_m,, m=1, …, M are error terms distributed as multivariate t-distribution (MVT), each with mean of zero, and variance-covariance matrix $\mathbf{\Sigma}$ as defined previously. The degree of freedom is set to four such that the variance of the multivariate t-distribution is twice of the standard normal distribution.

Two scenarios – identical and different regressors for the M-equations – are examined. Therefore, there are four sets of simulations in total. We generate 400 observations for each simulation and repeat the process up to 300 times. Since the values of X are fixed through repetitions, the empirical distributions of the parameter are used to provide standard errors for inference.

Identical Regressors

Regressors are identical across all four equations, including a constant term and an explanatory variable x, generated as two times a random uniform variable x=uniform( ). The variance-covariance matrix to generate the correlated error terms is $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.3 & -0.2 & 0.1 \\ 0.3 & 1 & 0.25 & 0.5 \\ -0.2 & 0.25 & 1 & 0.75 \\ 0.1 & 0.5 & 0.75 & 1 \end{bmatrix}$. And the values of beta are $\boldsymbol{\beta}$=[[3.45 -1],[5 -1.3],[0.8 -0.5],[-2.8 1.4]]´.

Different Regressors

Following Cappellari and Jenkins (2003), this study uses $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.25 & 0.5 & 0.75 \\ 0.25 & 1 & 0.75 & 0.5 \\ 0.5 & 0.75 & 1 & 0.75 \\ 0.75 & 0.5 & 0.75 & 1 \end{bmatrix}$, x_1=uniform( )-0.5, x_2=uniform( )+1/3, x_3=2*uniform( )+0.5, and x_4=0.5*uniform( )-1/3. The latent variables are generated as y1s=.5+4*x_1+ε_1, y2s=3+0.5*x_1-3*x_2+ ε_2, y3s = 1 - 2*x_1 + .4*x_2 - .75*x_3

+ ε_3, and y4s = -3.5 + 1*x_1 - .3*x_2 + 3*x_3 - .4*x_4 + ε_4, where the error terms are respectively generated from multivariate normal distribution and multivariate Stduent's t distribution with mean zero and variance-covariance $\mathbf{\Sigma}$. The binary outcome variables are then generated from above latent variable: y1 = (y1s>0), y2 = (y2s>0), y3 = (y3s>0), and y4 = (y4s>0).

Note for y4, this study uses -3.5 as the intercept instead of -6 in Cappellari and Jenkins (2003), to ensure a wider selection of mean values for the binary outcome variables. Table 2.1 reports the mean and cross correlation among the generated response variables under above four simulation scenarios.

2.3.2 Simulation Results

Parameter estimates and marginal effects, as well as model fit measured by log likelihood values are compared across multivariate binary discrete normal model and multivariate probit model (GHK simulation with 500 draws).

Tables 2.2 to 2.5 report parameter estimates for both models under each of the four simulation scenarios, respectively. A comparison across these two sets of parameter estimates shows that the pattern of signs and significance are similar between the two models, but by no means identical. The fits of the two models are almost identical – the 95% empirical confidence intervals of the difference in log-likelihood value between MBDN and MVP are not significantly different from zero (statistics not reported). In addition, Maximum Likelihood Estimation of the MVP model sometimes runs into problem because the combination of data set and initial values leads to a non-positive-definite covariance matrix for the GHK simulation. The MVP model experienced a failure rate of 28 percent, while the MBDN successfully ran for all 300 repetitions.

The average time required to run MVP model was about five times of that to run the MBDN model for the simulated data sets.

This paragraph now focuses on average marginal effects and some conditional expectations. Table 2.6 presents 16 sample average marginal effects for the case of identical regressors. Regardless whether the error terms are generated from MVN or MVT, the average marginal effects are very close in values under MBDN and MVP. Although some of them are statistically significant at 5% level, it's arguable from economic standpoint that they are not substantially different.

Tables 2.7 and 2.8 report average marginal effects for the case of different regressors with error term generated from MVN and MVT respectively. Note that because of the multivariate nature of the model, the marginal effects of x_2, x_3, and x_4 on y_1 are nonzero, even though they are not regressors for y_1. This is because there are indirect effects channeled through the correlation between y_1 and the other three response variables. Again, most of the marginal effects computed under MBDN and MVP are not statistically different from one another. For the few cases where the differences are statistically different from zero, the differences are relatively small in absolute values.

2.3.3 Correlations among Response Variables

As shown in Table 2.9, when the regressors are identical across equations, the MBDN estimates exactly match the observed correlation. This is because the first two moment equations are the maximum likelihood equations as proved in Section 2.3.3. When there are different regressors across equations, the first moment equation is still the maximum likelihood equation as long as there is a constant term. The second moment from the Multivariate Binary Discrete Normal

distribution is slightly different from the sample moment (Equation 2.18). Therefore in simulation cases 3 and 4, MBDN estimates of the correlation among response variables are slightly different from observed correlation. However, they are still much closer than the correlation among MVP residuals (correlation associated with the latent variables).

2.4 Applications to the Ketchup Brands Data

Using data provided by the James M. Kilts Center, University of Chicago Booth School of Business that was originally collected by the now-defunct ERIM division of A. C. Nielsen, this section examines the ketchup purchasing behavior of 1651 households in Sioux Falls, S.D. Data are from full calendar year 1986. The five brands studied represent more than 98% of all reported ketchup purchases. If the household is observed to purchase a given brand of ketchup at any time(s) during the year, then the response variable for that brand and household is coded one; otherwise it is coded zero. Table 2.9 summarizes the data and Table 2.10 suggests some merit to considering a multivariate approach.

Using the GHK simulator, a Multivariate Probit model was estimated. The simple model posits that the decision to purchase a given brand of ketchup is related to a polynomial in household size, whether the household lives in a single family house, the income of the household in $1000's, and the highest grade achieved by the head of the household. Results are reported in Table 2.11. Of particular interest is whether the response variables are correlated, and the test result reported in the table suggests this is the case.

Tables 2.12 and 2.15 report the exact maximum likelihood results using the multivariate binary discrete normal distribution and the multivariate probit model, respectively. Note that the

pattern of signs and significance are similar between the two models, but by no means identical. The fits of the two models are almost identical and since both models estimate the same number of parameters, application of an information criterion is unproductive

A focus on the marginal effects reveals almost an exact correspondence (at least to the first three decimal places) between the two models as shown in Table 2.14. Given that these marginal effects are simpler to derive from the multivariate binary discrete normal model and that estimation was achieved more than 20 times faster, this estimator appears to have merit in the analysis of large data sets. Finally, the implied correlations are presented in Table 2.15. The multivariate binary discrete normal reproduces the raw observed correlations. The Multivariate Probit's correlations are associated with the latent responses.

2.5 Conclusions and Discussions

A multivariate binary response model, termed as the Multivariate Binary Discrete Normal model, is obtained from the multivariate discrete normal distribution. The statistical model is a member of the quadratic exponential family and as such, under proper specification of conditional means, the maximum likelihood estimator possesses desirable properties even under distributional misspecification. Maximum likelihood estimation of the multivariate binary discrete normal model is straightforward because the normalizing factor for the joint probability mass function is obtained via the evaluation of 2^M support points for M binary response variables. The MBDN model nests the independent logit model. The MBDN estimates of the correlations among response variables coincide with observed ones when the regressors (including a constant term) are identical across equations. Lastly, marginal effects that count for the underlying property of multivariate model are much easier to derive and compute under the MBDN model.

Application of the statistical model to simulation data and to a well-known empirical data set suggests that the estimator can produce results with fits comparable to other estimators. In the empirical data, computation time is reduced by a factor of 20 relative to a MVP model.

Table 2.1 Summary of Descriptive Statistics of the Generated Response Variables

Mean (average over 300 repetitions)					
		y_1	y_2	y_3	y_4
Case 1: Identical regressors + MVN		0.9250	0.9775	0.5400	0.4750
Case 2: Identical regressors + MVT		0.9150	0.9600	0.5475	0.4825
Case 3: Different regressors + MVN		0.6194	0.6394	0.7783	0.1537
Case 4: Different regressors + MVT		0.6750	0.6900	0.8150	0.2075
Correlation (the last data set for example)					
		y_1	y_2	y_3	y_4
Case 1: Identical regressors + MVN	y_1		0.3177	0.0683	-0.2651
	y_2			0.1221	-0.1965
	y_3				-0.0912
Case 2: Identical regressors + MVT	y_1		0.2100	0.0452	-0.2434
	y_2			0.1335	-0.0978
	y_3				-0.0374
Case 3: Different regressors + MVN	y_1		0.0646	-0.1206	0.1533
	y_2			0.1271	0.1168
	y_3				0.0382
Case 4: Different regressors + MVT	y_1		0.0964	-0.0415	0.2072
	y_2			0.1480	0.1282
	y_3				0.0458

Note: N=400 for each simulated data set

Table 2.2 Parameter Estimates: Identical Regressor and MVN

		Btrue	MVP [a]	95% Empirical CI		MBDN [b]	95% Empirical CI	
y_1	Intercept	3.45	**3.4872***	2.8074	4.5319	**9.4064***	4.1471	13.5649
	X	-1	**-1.0083***	-1.3815	-0.7545	**-2.4984***	-3.7249	-1.1026
y_2	Intercept	5	**5.3019***	3.8622	8.0935	**10.4651***	6.6867	15.4808
	X	-1.3	**-1.3913***	-2.3174	-0.8547	**-2.5537***	-4.2306	-1.2714
y_3	Intercept	0.8	**0.7975***	0.5636	1.0205	**4.7663**	-1.0632	10.8475
	X	-0.5	**-0.4971***	-0.6141	-0.3629	**-1.0765**	-2.9187	0.7417
y_4	Intercept	-2.8	**-2.8600***	-3.3343	-2.4073	**4.8362**	-1.4653	10.7174
	X	1.4	**1.4277***	1.1933	1.6788	**-0.7035**	-2.7888	1.4820
	σ_{12}	0.3	**0.2918**	-0.0309	0.6248	**0.4205**	-0.3168	0.7886
	σ_{13}	-0.2	**-0.1880**	-0.4045	0.0258	**0.0853**	-0.5410	0.6073
	σ_{14}	0.1	**0.1140**	-0.1602	0.3792	**0.2713**	-0.4693	0.7126
	σ_{23}	0.25	**0.2407**	-0.0723	0.7234	**0.6082***	0.0881	0.8879
	σ_{24}	0.5	**0.4626***	0.0591	0.7674	**0.6992***	0.2320	0.8605
	σ_{34}	0.75	**0.7480***	0.5878	0.8889	**0.7925***	0.6731	0.8685
Log likelihood			**-512.75**	-551.03	-481.12	**-517.37**	-557.8	-487.65
Repetitions			240			240		

[a] Average estimates over all repetitions under the MVP model
[b] Average estimates over all repetitions under the MBDN model
* Estimates are significantly different from zero at 5% level (based on empirical confidence interval)

Table 2.3 Parameter Estimates: Identical Regressor and MVT

		Btrue	MVP [a]	95% Empirical CI		MBDN [b]	95% Empirical CI	
y_1	Intercept	3.45	**2.5710***	2.0707	3.3645	**6.7215***	4.3643	8.7547
	X	-1	**-0.7062***	-0.9978	-0.5091	**-1.6343***	-2.3137	-1.0343
y_2	Intercept	5	**3.2118***	2.4947	4.4541	**6.6968***	4.2842	9.2862
	X	-1.3	**-0.7449***	-1.1824	-0.4509	**-1.4430***	-2.316	-0.6201
y_3	Intercept	0.8	**0.7380***	0.4996	0.9566	**2.2552**	-1.4681	5.3254
	X	-0.5	**-0.4588***	-0.5926	-0.33	**-0.4275**	-1.4218	0.6292
y_4	Intercept	-2.8	**-2.3154***	-2.7663	-1.9508	**2.6129**	-0.8039	5.3878
	X	1.4	**1.1614***	0.9587	1.3864	**-0.1403**	-1.0875	1.0451
	σ_{12}	0.3	**0.3616***	0.0663	0.6306	**0.5368**	-0.1302	0.7824
	σ_{13}	-0.2	**-0.2013**	-0.4455	0.0076	**0.0430**	-0.5111	0.4898
	σ_{14}	0.1	**0.0874**	-0.1649	0.3658	**0.2933**	-0.3416	0.6791
	σ_{23}	0.25	**0.1768**	-0.1265	0.4455	**0.5035***	0.0386	0.7662
	σ_{24}	0.5	**0.4861***	0.1631	0.7553	**0.7030***	0.3906	0.8324
	σ_{34}	0.75	**0.7066***	0.5292	0.8582	**0.7646***	0.6136	0.8431
Log likelihood			**-590.28**	-630.3	-549.73	**-592.28**	-633.56	-554.36
Repetitions			200			200		

* Estimates are significantly different from zero at 5% level (based on empirical confidence interval)

Table 2.4 Parameter Estimates: Different Regressors and MVN

		Btrue	MVP	95% Empirical CI		MBDN	95% Empirical CI	
y_1	Intercept	0.5	**0.5137***	0.3661	0.6886	**1.1455***	0.6944	1.7673
	X_1	4	**4.0676***	3.4671	4.8008	**5.0473***	2.9378	7.5436
y_2	Intercept	3	**3.0192***	2.4951	3.5958	**2.8278***	1.8504	5.5827
	X_1	0.5	**0.4941***	0.0655	0.9527	**2.2065**	-0.7852	4.2439
	X_2	-3	**-3.0194***	-3.6123	-2.4614	**-1.9964***	-5.1332	-1.1018
y_3	Intercept	1	**1.0120***	0.5963	1.4785	**1.5551***	0.3053	2.6503
	X_1	-2	**-2.0510***	-2.6518	-1.4927	**1.5753**	-0.9511	3.5665
	X_2	0.4	**0.3988**	-0.071	0.9236	**-0.4046**	-3.012	0.2812
	X_3	-0.75	**-0.7532***	-1.0262	-0.5097	**-0.0485**	-0.4729	1.5728
y_4	Intercept	-3.5	**-3.6124***	-4.7337	-2.6899	**-0.7686**	-6.2506	0.7897
	X_1	1	**1.0218***	0.3132	1.7595	**2.8142***	0.3301	4.8789
	X_2	-0.3	**-0.3209**	-0.9396	0.3787	**-0.3630**	-1.7262	0.1221
	X_3	3	**3.1054***	2.5166	3.7969	**1.5117***	0.5885	5.8239
	X_4	-0.4	**-0.4174**	-1.7784	0.9407	**-0.1112**	-0.9933	0.7121
	σ_{12}	0.25	**0.2346***	0.0033	0.4331	**0.3500**	-0.0172	0.4316
	σ_{13}	0.5	**0.5099***	0.3031	0.743	**0.3855***	0.0530	0.4428
	σ_{14}	0.75	**0.7287***	0.4930	0.8937	**0.3634***	0.0003	0.4479
	σ_{23}	0.75	**0.7462***	0.6001	0.8622	**0.4373***	0.3658	0.4575
	σ_{24}	0.5	**0.5046***	0.2380	0.7251	**0.3689***	0.0715	0.4471
	σ_{34}	0.75	**0.7597***	0.5691	0.9231	**0.4118***	0.2281	0.453
Log likelihood			**-559.01**	-595.98	-523.34	**-585.86**	-625.73	-547.91
Repetitions			**187**			187		

* Estimates are significantly different from zero at 5% level (based on empirical confidence interval)

Table 2.5 Comparison of MBDN and MVP Estimates: Different Regressors and MVT

		Btrue	MVP	95% Empirical CI		MBDN	95% Empirical CI	
y_1	Intercept	0.5	**0.4100***	0.2694	0.5736	**0.9077***	0.4835	1.4443
	X_1	4	**3.3900***	2.8533	4.0514	**4.1058***	2.5171	5.9076
y_2	Intercept	3	**2.6088***	2.1324	3.1098	**2.2563***	1.6002	4.289
	X_1	0.5	**0.4150**	-0.0209	0.8798	**2.0941***	0.3818	3.7578
	X_2	-3	**-2.6142***	-3.1427	-2.0951	**-1.5642***	-3.6269	-1.0036
y_3	Intercept	1	**0.8635***	0.4451	1.3802	**1.3561***	0.7897	2.4262
	X_1	-2	**-1.6433***	-2.2285	-1.0481	**1.4665**	-0.2331	3.2123
	X_2	0.4	**0.3303**	-0.1877	0.8552	**-0.1847**	-1.626	0.277
	X_3	-0.75	**-0.6389***	-0.888	-0.3788	**-0.2147**	-0.4326	0.6999
y_4	Intercept	-3.5	**-2.6432***	-3.5946	-1.847	**-0.2316**	-3.6073	0.5882
	X_1	1	**0.8471***	0.2273	1.582	**2.5821***	0.7679	4.2904
	X_2	-0.3	**-0.2264**	-0.8476	0.3595	**-0.2537**	-1.2418	0.1273
	X_3	3	**2.2410***	1.691	3.1243	**0.8858***	0.5043	4.2019
	X_4	-0.4	**-0.3069**	-1.4722	0.8466	**-0.1073**	-0.8153	0.3915
	σ_{12}	0.25	**0.2396***	0.0358	0.428	**0.6918***	0.3062	0.7952
	σ_{13}	0.5	**0.4825***	0.2780	0.6846	**0.7565***	0.5736	0.8347
	σ_{14}	0.75	**0.7159***	0.4967	0.904	**0.7590 ***	0.0877	0.8511
	σ_{23}	0.75	**0.7353***	0.6006	0.8674	**0.8373***	0.7481	0.8774
	σ_{24}	0.5	**0.4697***	0.2319	0.7204	**0.7431***	0.1735	0.8398
	σ_{34}	0.75	**0.7171***	0.5020	0.8795	**0.7883***	0.4023	0.8576
Log likelihood			**-626.18**	-662.68	-587.71	**-647.51**	-686.94	-603.9
Repetitions			240			240		

* Significant at 5% level (based on empirical confidence interval)

Table 2.6 Sample Average Marginal Effects: Identical Regressor

Marginal Effect of X upon:	Error Term ~ MVN						Error Term ~ MVT					
	MBDN		MVP		MBDN -MVP		MBDN		MVP		MBDN -MVP	
E(y1)	-0.1609	**	-0.1442	**	-0.0167	*	-0.1407	**	-0.1278	**	-0.0129	**
E(y2)	-0.1059	**	-0.0918	**	-0.0141	*	-0.0906	**	-0.0771	**	-0.0134	*
E(y3)	-0.2070	**	-0.1739	**	-0.0331	**	-0.1905	**	-0.1628	**	-0.0277	**
E(y4)	0.2759	**	0.2822	**	-0.0063		0.2666	**	0.2720	**	-0.0054	
E(y1|y2=1)	-0.1406	**	-0.1265	**	-0.0140	*	-0.1152	**	-0.1091	**	-0.0060	
E(y1|y3=1)	-0.1889	**	-0.1752	**	-0.0137		-0.1740	**	-0.1607	**	-0.0133	
E(y1|y4=1)	-0.1594	**	-0.1436	**	-0.0158		-0.1383	**	-0.1285	**	-0.0098	
E(y2|y1=1)	-0.0762	**	-0.0679	**	-0.0083		-0.0555	**	-0.0522	**	-0.0033	
E(y2|y3=1)	-0.0711	**	-0.0588	**	-0.0123		-0.0758	**	-0.0549	**	-0.0209	**
E(y2|y4=1)	-0.1018	**	-0.0897	**	-0.0121	*	-0.0838	**	-0.0722	**	-0.0116	*
E(y3|y1=1)	-0.2171	**	-0.1854	**	-0.0317	**	-0.2028	**	-0.1748	**	-0.0280	**
E(y3|y2=1)	-0.1991	**	-0.1649	**	-0.0343	**	-0.1868	**	-0.1566	**	-0.0301	**
E(y3|y4=1)	-0.2388	**	-0.2344	**	-0.0043		-0.2269	**	-0.2270	**	0.0001	
E(y4|y2=1)	0.2794	**	0.2833	**	-0.0039		0.2715	**	0.2732	**	-0.0017	
E(y4|y3=1)	0.2822	**	0.2858	**	-0.0036		0.2765	**	0.2793	**	-0.0029	
E(y4|y4=1)	0.2863	**	0.2878	**	-0.0015		0.2821	**	0.2844	**	-0.0023	

** Significant at 5% level (based on empirical confidence interval);
* Significant at 10% level.

Table 2.7 Sample Average Marginal Effects: Different Regressors and MVN

Marginal Effect of X_i upon:	**X1**			**X2**			**X3**			**X4**		
	MBDN	**MVP**	**Diff[1]**	**MBDN**	**MVP**	**Diff**	**MBDN**	**MVP**	**Diff**	**MBDN**	**MVP**	**Diff**
E(y1)	0.927**	0.926**	0.001	0.096*	0.001	0.095	-0.066	0.000	-0.067	0.012	-0.000	0.013
E(y2)	0.116*	0.132**	-0.016	-0.773**	-0.813**	0.040	-0.051	0.000	-0.050	0.012	0.000	0.011
E(y3)	-0.524**	-0.484**	-0.041	0.188**	0.093*	0.095**	-0.22**	-0.178**	-0.042	0.013	-0.000	0.013
E(y4)	0.105**	0.127**	-0.022	0.030	-0.041	0.070	0.294**	0.389**	-0.095	-0.030	-0.053	0.023
E(y1|y2=1)	0.920**	0.906**	0.014	0.136	0.082**	0.053	-0.044	0.000	-0.044	0.010	0.000	0.010
E(y1|y3=1)	0.934**	0.931**	0.004	0.064	-0.012	0.076	-0.027	0.026**	-0.053	0.009	0.000	0.009
E(y1|y4=1)	0.821**	0.243	0.578*	0.071	0.137	-0.066	-0.078	-0.110*	0.032	0.014	0.032	-0.020
E(y2|y1=1)	0.061	0.010	0.051	-0.77**	-0.787**	0.017	-0.020	0.000	-0.020	0.008	-0.000	0.008
E(y2|y3=1)	0.275**	0.270**	0.006	-0.763**	-0.785**	0.022	0.028	0.053**	-0.025	0.006	-0.000	0.006
E(y2|y4=1)	0.135**	-0.082	0.216	-0.658**	-0.857	0.199	-0.046	-0.200	0.154	0.011	-0.052	0.060
E(y3|y1=1)	-0.548**	-0.514**	-0.033	0.150**	0.075	0.075**	-0.168**	-0.143**	-0.025	0.009	-0.000	0.009
E(y3|y2=1)	-0.410**	-0.372**	-0.038**	0.288**	0.239**	0.049*	-0.144**	-0.127**	-0.018	0.007	-0.000	0.007
E(y3|y4=1)	-0.339**	-0.185**	-0.153**	0.121**	0.042**	0.079**	-0.165**	-0.164**	0.000	0.011	0.015	-0.002
E(y4|y2=1)	0.003	-0.058	0.060	0.016	-0.045	0.061	0.367**	0.462**	-0.095**	-0.039	-0.062	0.023
E(y4|y3=1)	0.131**	0.124**	0.006	0.123	0.058	0.065	0.352**	0.433**	-0.082**	-0.037	-0.058	0.021
E(y4|y4=1)	0.222**	0.236**	-0.015	-0.002	-0.063	0.061	0.375**	0.454**	-0.079**	-0.037	-0.056	0.020

[1]Diff is computed as the difference between MBDN estimates and MVP estimates;
** Significant at 5% level (based on empirical confidence interval);
* Significant at 10% level.

Table 2.8 Sample Average Marginal Effects: Different Regressors and MVT

Marginal Effect of X_i upon:	**X1**			**X2**			**X3**			**X4**		
	MBDN	**MVP**	**Diff[1]**	**MBDN**	**MVP**	**Diff**	**MBDN**	**MVP**	**Diff**	**MBDN**	**MVP**	**Diff**
E(y1)	0.883**	0.879**	0.004	0.105**	-0.001	0.106**	-0.081*	0.001	-0.082**	0.014	0.000	0.014
E(y2)	0.104	0.120*	-0.015	-0.717**	-0.762**	0.045	-0.054**	0.001	-0.055**	0.012	-0.000	0.012
E(y3)	-0.482**	-0.436**	-0.046**	0.189**	0.087	0.102**	-0.219**	-0.167**	-0.052**	0.014	-0.000	0.014
E(y4)	0.105**	0.133**	-0.028	0.041	-0.036	0.077**	0.263**	0.351**	-0.088*	-0.030	-0.049	0.019
E(y1|y2=1)	0.874**	0.852**	0.022	0.154**	0.084**	0.070	-0.052*	0.000	-0.052*	0.011	0.000	0.011
E(y1|y3=1)	0.897**	0.886**	0.011	0.068	-0.014	0.082*	-0.031	0.027**	-0.059*	0.010	0.000	0.010
E(y1|y4=1)	0.719**	0.327*	0.393*	0.071	-0.026	0.097	-0.092*	-0.154	0.062	0.015	0.034	-0.020
E(y2|y1=1)	0.039	-0.002	0.041	-0.715**	-0.734**	0.019	-0.015	0.000	-0.015	0.008	0.000	0.008
E(y2|y3=1)	0.265**	0.252**	0.014	-0.716**	-0.736**	0.020	0.033	0.054**	-0.022	0.006	0.000	0.006
E(y2|y4=1)	0.114*	0.165*	-0.052	-0.592**	-0.793	0.201	-0.052	0.179	-0.231	0.012	0.169	-0.157
E(y3|y1=1)	-0.519**	-0.487**	-0.032	0.145**	0.070	0.074**	-0.157**	-0.135**	-0.022	0.008	0.000	0.008
E(y3|y2=1)	-0.370**	-0.337**	-0.033*	0.294**	0.244**	0.050**	-0.137**	-0.119**	-0.018	0.006	0.000	0.006
E(y3|y4=1)	-0.313**	-0.187**	-0.127**	0.120**	0.041	0.079**	-0.157**	-0.150**	-0.007	0.011	0.012	-0.001
E(y4|y2=1)	-0.033	-0.071	0.037	0.024	-0.046	0.070	0.354**	0.439**	-0.085**	-0.041	-0.062	0.021
E(y4|y3=1)	0.132**	0.136**	-0.004	0.150**	0.070	0.080	0.330**	0.400**	-0.070**	-0.037	-0.056	0.019
E(y4|y4=1)	0.217**	0.250**	-0.033	0.011	-0.061	0.072*	0.352**	0.431**	-0.080**	-0.037	-0.055	0.018

[1]Diff is computed as the difference between MBDN estimates and MVP estimates;
** Significant at 5% level (based on empirical confidence interval);
* Significant at 10% level.

Table 2.9 Correlation Estimates among Response Variables

Case 1. Identical Regressors + MVN

	MBDN				MVP Residuals		
	y_2	y_3	y_4		y_2	y_3	y_4
y_1	0.3177	0.0683	-0.2651	y_1	0.9084	0.0862	-0.6396
y_2		0.1221	-0.1965	y_2		0.1340	-0.6825
y_3			-0.0912	y_3			0.0945

MBDN estimates are identical to sample correlation as shown in Table 1

Case 2. Identical Regressors + MVT

	MBDN				MVP Residuals		
	y_2	y_3	y_4		y_2	y_3	y_4
y_1	0.2100	0.0452	-0.2434	y_1	0.6530	0.0260	-0.4079
y_2		0.1335	-0.0978	y_2		0.1367	-0.4491
y_3			-0.0374	y_3			0.1536

MBDN estimates are identical to sample correlation as shown in Table 1

Case 3. Different Regressors + MVN

	MBDN				MVP Residuals		
		y_2	y_3	y_4	y_2	y_3	y_4
y_1	0.0514	-0.1060	0.1319	y_1	0.0836	-0.1493	0.1517
y_2		0.1266	0.1077	y_2		0.1123	0.1342
y_3			0.0414	y_3			-0.2231

MBDN estimates are closer to the sample observation

Case 4. Different Regressors + MVT

	MBDN				MVP Residuals		
		y_2	y_3	y_4	y_2	y_3	y_4
y_1	0.1001	-0.0221	0.1790	y_1	0.1680	0.0032	0.1597
y_2		0.1456	0.1416	y_2		0.1658	0.1497
y3			0.0519	y3			-0.1123

MBDN estimates are identical to sample correlation as shown in Table 1

Note: Estimates from the last data set for illustration.

Table 2.10 Counts of Households Purchasing Ketchup (216 Households Make No Purchases)

Combination of Brand(s)	With **Brand1**	**Brand2**	**Brand3**	**Brand4**	**Brand5**
1	**1258**				
2	343	**426**			
3	465	223	**555**		
4	96	56	76	**123**	
5	134	81	102	19	**173**
2&3	198				
2&4	51				
2&5	69				
3&4	68	43			
3&5	83	56			
4&5	16	12	16		
2&3&4	41				
2&3&5	50				
3&4&5	14	12			
2&3&4&5	11				

Table 2.11 Correlations of Observed Binary Responses: Ketchup Brands

	Brand 1	**Brand 2**	**Brand 3**	**Brand 4**	**Brand 5**
Brand 1	1	0.0598	0.1268	0.0123	0.0101
Brand 2		1	0.2338	0.1279	0.1644
Brand 3			1	0.1692	0.1836
Brand 4				1	0.046
Brand 5					1

Table 2.12 Ketchup Brands Maximum Likelihood Results—MVP (GHK Simulator)
Log Likelihood = -3607.05

Brand1	**Coefficient**	**Std. Error**	**z-Value**	**Brand2**	Coefficient	Std. Error	z-Value
Intercept	-0.6103	0.1821	-3.3511	Intercept	-1.448	0.192	-7.5417
Size	0.7284	0.1137	6.4051	Size	0.3729	0.1059	3.5204
Size^2	-0.0649	0.0172	-3.7623	Size^2	-0.0304	0.0151	-2.0168
House	0.1357	0.0992	1.368	House	0.3055	0.111	2.7519
Income	0.0002	0.0025	0.0753	Income	-0.0040	0.0023	-1.7478
Educ	-0.0192	0.0188	-1.0181	Educ	-0.0174	0.018	-0.9648
Brand3	Coefficient	Std. Error	z-value	**Brand4**	Coefficient	Std. Error	z-value
Intercept	-1.3703	0.1803	-7.6001	Intercept	-1.2189	0.2273	-5.3622
Size	0.524	0.1013	5.1709	Size	0.0738	0.1440	0.5124
Size^2	-0.0424	0.0145	-2.9304	Size^2	0.0045	0.0197	0.2262
House	-0.0246	0.1005	-0.2448	House	-0.1044	0.1353	-0.7715
Income	-0.0058	0.0024	-2.4558	Income	-0.0095	0.0051	-1.8599
Educ	0.0064	0.0171	0.3761	Educ	-0.0200	0.0260	-0.7675
Brand5	Coefficient	Std. Error	z-Value				
Intercept	-2.0977	0.2529	-8.2932				
Size	0.3852	0.1335	2.8858				
Size^2	-0.0307	0.0183	-1.6753				
House	0.1651	0.1435	1.150				
Income	-0.0044	0.0030	-1.4577				
Educ	-0.0004	0.0231	-0.0182				
r_{12}	-0.0052	0.0473	-0.1109	r_{24}	0.2840	0.0580	4.8963
r_{13}	0.1017	0.0456	2.2324	r_{25}	0.2967	0.0527	5.6274
r_{14}	-0.0157	0.0637	-0.2472	r_{34}	0.3720	0.0543	6.8564
r_{15}	-0.1004	0.0592	-1.6963	r_{35}	0.3178	0.0500	6.3553
r_{23}	0.3403	0.0408	8.3504	r_{45}	0.1155	0.0729	1.5850

LR Test of H_0: $\rho_{ij}=0$ all i<j

X^2 = 172.3 with 10 df.
p = 0.0000

Table 2.13 Ketchup Brands Maximum Likelihood Results—MBDN
Log Likelihood = -3606.03

Brand1	Coefficient	Std. Error	z-Value	**Brand2**	Coefficient	Std. Error	z-Value
Intercept	0.0566	1.384	0.0409	Intercept	-4.7266	0.5202	-9.0869
Size	1.0734	0.3984	2.694	Size	1.2382	0.3090	4.0065
Size^2	-0.0911	0.0425	-2.1451	Size^2	-0.0928	0.0382	-2.4323
House	0.1465	0.1901	0.7706	House	0.4604	0.3059	1.5051
Income	0.002	0.0086	0.2353	Income	-0.0259	0.0091	-2.8383
Educ	-0.0344	0.0344	-1.000	Educ	-0.0346	0.0509	-0.6790
Brand3	Coefficient	Std. Error	z-Value	**Brand4**	Coefficient	Std. Error	z-Value
Intercept	-4.7816	0.5160	-9.2664	Intercept	-4.3343	0.5511	-7.8642
Size	1.4267	0.3110	4.5883	Size	1.0388	0.3588	2.895
Size^2	-0.1075	0.0384	-2.7967	Size^2	-0.0700	0.0434	-1.6129
House	0.2670	0.3027	0.8822	House	0.1505	0.3214	0.4683
Income	-0.0279	0.0096	-2.8955	Income	-0.0308	0.0122	-2.5302
Educ	-0.0250	0.0518	-0.4826	Educ	-0.0402	0.0593	-0.6776
Brand5	Coefficient	Std. Error	z-Value				
Intercept	-4.951	0.5848	-8.4662				
Size	1.085	0.3269	3.3192				
Size^2	-0.0811	0.0410	-1.9799				
House	0.3868	0.3379	1.1444				
Income	-0.0236	0.0090	-2.6087				
Educ	-0.0144	0.0544	-0.2641				
s_{15}	-0.1864	0.2088	-0.8926	**LR Test of H_0: $\sigma_{ij}=0$ all i<j**			
s_{23}	0.7088	0.0232	30.5638				
s_{24}	0.6627	0.0448	14.7817	X^2 = 179.8 with 10 df.			
s_{25}	0.6461	0.0482	13.3995	p = 0.0000			
s_{34}	0.7124	0.0367	19.4051				
s_{35}	0.6472	0.0529	12.2345				
s_{45}	0.5278	0.1044	5.0575				

Table 2.14 Comparison of Average Marginal Effects for Ketchup Brands Choice

	Brand 1	**Brand 2**	**Brand 3**	**Brand 4**	**Brand 5**
MVP—Size +1	0.08802	0.053	0.08492	0.015345	0.03096
MBDN—Size +1	0.08874	0.05292	0.08493	0.016987	0.03114
MVP—Income+1[a]	0.00006	-0.00124	-0.00199	-0.001299	-0.00075
MBDN—Income+1	0.00012	-0.00133	-0.00219	-0.001563	-0.00079

[a]Income in $1,000's

Table 2.15 Ketchup Brands Choice Correlations Implied by Models—MDBN Coincides with Observed Correlations

Brand	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
1-MVP	1	-0.0052	0.1017	-0.0157	-0.1004
1-MBDN		0.0598	0.1268	0.0123	0.0101
2-MVP		1	0.3403	0.2840	0.2967
2-MBDN			0.2338	0.1279	0.1644
3-MVP			1	0.3720	0.3178
3-MBDN				0.1692	0.1836
4-MVP				1	0.1155
4-MBDN					0.0460
5-MVP					1
5-MBDN					

CHAPTER 3

POLISH HOUSEHOLD CONSUMPTION OF TOBACCO AND ALCOHOL: A CENSORED SYSTEM

The addictive nature of tobacco and alcohol suggests that it is more appropriate to model their consumption as a system. The approach adopted in this chapter expands Heckman's sample selection model into a censored system (or multivariate sample selection model) to analyze household tobacco and alcohol consumption. The estimation uses pooled cross-sectional data of 77,043 observations from Polish Household Survey data in the period of 2005 to 2008 and applies full information maximum likelihood estimation. Empirical investigation indicates that the decisions to smoke and drink as well as their expenditure levels are positively correlated. The empirical investigation involves the effects of demographic and socio-economic factors as well as outmigration, a special issue in Poland, on the consumption decisions and expenditures on tobacco and alcohol. Findings provide insights for the reduction and prevention of tobacco and alcohol use.

3.1 Introduction

Tobacco use is one of the main risk factors for a number of chronic diseases, including cancer, lung diseases, and cardiovascular diseases. The World Health Organization has estimated that tobacco use is currently responsible for the deaths of about six million people across the world each year (WHO 2016c). Institutions and governments are taking steps to reduce tobacco use (e.g., the World Health Organization Framework Convention on Tobacco Control, WHO FCTC).

Similarly, the consumption of alcohol carries a risk of adverse health and social consequences related to its inebriating, toxic, and dependence-producing properties (WHO 2016d).

Tobacco and alcohol-related harm are two high-priority public health challenges facing Europe. In Poland, 29.4% of the population is estimated to use smoking tobacco in 2013 (WHO 2016e). Per capita (15 years and older) alcohol consumption in Poland is steadily increasing during the period 2000–2010 (WHO 2016f). With the growing attention to tobacco- and alcohol-related social and health problems and public policy campaigns against tobacco and alcohol use, an analysis of the determinants of household tobacco and alcohol consumption remains important. This study takes advantage of second-hand survey data collected from a household panel by Poland's Main Statistical Office (GUS) that is not publicly available.

Earlier studies on individual tobacco and alcohol use have identified a variety of demographic and socio-economic factors as consumption determinants, including income, education level, age, gender, region of residence, and employment status (e.g., Blaylock and Blisard 1992; Jones and Labeaga 2003; Yen 2005). Income is one of the most commonly used variables in studies of cigarettes and alcohol (Yen 2005). Individuals with a higher educational attainment level may be more aware of the risks of tobacco and alcohol consumption than those with less education. Urban residence, compared to rural residence, and employment status may reflect different lifestyles and economic wellbeing. Age is relevant as a previous study suggests a life-cycle pattern for smoking (Freeth 1998) and such a pattern is likely for alcohol consumption. Furthermore, WHO reports on tobacco and alcohol use clearly reveal different patterns for female drinkers/smokers compared to their male counterparts (WHO 2016c, 2016d).

Household food consumption literature suggests household size and structure also plays a role in household consumption decisions. The presence of children generally is associated with

healthier food choices. Also, the presence of elders may also indicate differences in consumption patterns. This study, therefore, includes household size as an explanatory variable. This measurement is further broken down into the number of adults as well as the presence and numbers of children and elders. Specifically, the presence of children and elders are assumed to affect the participation decisions of whether to buy tobacco or alcohol, while the number of adult members is assumed to affect the consumed amounts. The number of adults is further broken down into two categories, those under 60 years and those age 60 or above (elders). The purpose of this break down is to capture possibly different consumption patterns among these two groups of people.

A special factor in Poland is worker migration and depopulation, especially after Poland's accession to the EU in 2004, coupled with free job market entry to other EU countries. Migration leads to changes in population structure and exposure to different lifestyle and cultural values, which in return contributes to different consumption features. Previous studies focused on the dampening effect of depopulation on economic growth; however, less attention has been paid to the dietary welfare of people living in the depopulating regions at a micro or household level. This study investigates determinants of household expenditure on tobacco and alcohol, with special attention paid to the effect of depopulation associated with worker migration, a current issue in Poland. Additionally, the study takes into account the effects of migration to other countries. The study applies quantitative methods to generate measurable effects of individual explanatory factors.

The remainder of this chapter is organized as follows: Section 3.2 describes the methodology, including economic theory and econometric modeling. Section 3.3 introduces data sources and variable definitions. Section 3.4 reports estimation results. Finally, Section 3.5 concludes with discussion.

3.2 Modeling Approach

3.2.1 Economic Theory

A qualitative choice model based on random utility maximization developed by McFadden (1980) provides the theoretical foundation for model specification. An empirical model is derived by extending the discrete choice model (Pudney, 1989). A household maximizes the random utility function subject to a budget constraint. The household random utility function is given by:

(3.1) $$V(y, q; \boldsymbol{w}) = d \cdot U(y, q; \boldsymbol{w}) + (1 - d) \cdot U^*(q; \boldsymbol{w})$$

where U is the utility for buyers and U* for non-purchasers, y is the quantity of a commodity with price p, q is a composite commodity for other goods with price normalized to one, w is a vector of demographic variables, and d is a binary variable that equals one if the household buys the commodity and zero otherwise.

Assume the outcome for tobacco and alcohol consumption, the participation decision, is generated by a binary choice structure:

(3.2) $$d = 1 \; if \; \boldsymbol{z}'\boldsymbol{\alpha} + u > 0$$

$$= 0 \; if \; \boldsymbol{z}'\boldsymbol{\alpha} + u \leq 0,$$

where **z** and **α** are vectors of variables and parameters affecting binary purchase decision, and u is a random error.

In cross-sectional demand modeling, zero observations are often treated as the result of economic non-consumption (i.e., a corner solution). In some cases, however, zero purchase might be caused by behavioral factors other than prices. Because y does not enter the purchasers' utility function $U^*(q; \boldsymbol{w})$ as described in Equation (3.1) and p > 0, the optimal level is y = 0 for a non-consumer. This optimal zero purchase could be a corner solution or the result of opting out of the

market. For a buyer, the optimal level of y results from a solution to the constrained utility maximization problem with a fixed budget I:

(3.3) $\max_{y,q} \{U(y, q; \boldsymbol{w}) | \, py + q = I\}$.

Assume that the utility function $U(y, q; \boldsymbol{w})$ is regular strictly quasi-concave and has positive first partial derivatives with respect to y and q. Furthermore, assume an interior solution for y and q. Then, solving Equation (3.3) yields the notional (latent) demand for alcohol or tobacco, y*. Denote as $\boldsymbol{x}$ the vector of income and demographic variables (with corresponding parameter vector $\boldsymbol{\beta}$) affecting the quantity demanded.

Further, assume latent quantity y* is expressed by the lognormal distribution, which accommodates right-skewness and ensures positive purchase amount:

(3.4) $y^* = \boldsymbol{x}'\boldsymbol{\beta} + v$,

where $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are variables and corresponding parameters affecting quantity decision and v is a random error.

3.2.2 Econometric Modeling

The occurrence of excessive percentage of zeros in micro-data sets mandates a proper treatment for the censoring of the dependent variables. Such zero observations may occur for three main reasons: infrequency of purchase in survey data with short recording periods, some individuals withdraw from the market for various reasons, and economic non-consumption under current price and individual income.

The particular interpretation given to zero observations can have a crucial bearing on the estimation approach adopted (Madden 2008). Various modeling structures are proposed in existing

literature to accommodate the censored data, including the Tobit model, hurdle model, two-part model, and Heckman's sample selection model. More recent developments feature a sample selection system or censored system in the sense of multiple-goods decisions, which allows correlation within and/or across participation decisions and intensity decisions among multiple goods. Such feature is important for studying the consumption of closely related products, such as the consumption of tobacco and alcohol. A number of censored-system estimation procedures have existed in the literature. These include maximum-likelihood estimators of Amemiya (1995), Wales and Woodland (1983), and Lee and Pitt (1986), and two-step estimators of Heien and Wessells (1990), Shonkwiler and Yen (1999), and Perali and Chavas (2000), as well as an extended full system approach of Stewart and Yen (2004), and Yen (2005).

Due to the additive nature as well as observed culture of drinking and smoking, it is more appropriate to model their consumption as a system (Pierani and Tiezzi 2009). This study uses a censored system which specifies a set of level equations, each exclusively subject to a binary selection rule, and which accommodates error correlations among all equations.

To facilitate the presentation of models, the binary choice rules and level equations, described by Equations (3.2) and (3.4), respectively, are re-written in a system. Then, each outcome variable y_i is governed by a binary selection rule of whether to consume as follows (observation subscription omitted):

$$(3.5) \qquad \log(y_i) = \mathbf{x}'\boldsymbol{\beta}_\mathbf{i} + v_i \quad \text{if } \mathbf{z}'\boldsymbol{\alpha}_\mathbf{i} + u_i > 0$$

$$y_i = 0 \quad \text{if } \mathbf{z}'\boldsymbol{\alpha}_\mathbf{i} + u_i \leq 0 \ , \quad i = 1,2$$

where $\mathbf{z}$ and $\mathbf{x}$ are vectors affecting binary purchase decision and level decision, respectively; $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of parameters; and $\boldsymbol{u_i}$ and $\boldsymbol{v_i}$ are random errors in the participation and level equation, respectively.

To facilitate presentation of the log likelihood functions, define diagonal $S = \mathrm{diag}(\sigma_1, \sigma_2)$ as standard deviation of $\boldsymbol{v}$. Let $\mathbf{R_{uu}} = [\rho_{ij}^{uu}]$, $\mathbf{R_{vu}} = [\rho_{ij}^{vu}]$, and $\mathbf{R}_{\boldsymbol{vv}} = [\rho_{ij}^{vv}]$ be 2 x 2 correlation matrices among elements of $\boldsymbol{u}$ and $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{u}$, and $\boldsymbol{v}$ and $\boldsymbol{v}$, respectively.

The censored system, which allows error correlations among all equations, assumes the concatenated error vector $[\boldsymbol{u}', \boldsymbol{v}']' \equiv [u_1, u_2, v_1, v_2]'$ is distribted as 4-variate normal with zero mean and covariance matrix:

(3.6) $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$,

where $\Sigma_{11} = \mathrm{E}(\boldsymbol{uu}') = \mathbf{R}_{uu}$, $\Sigma_{21} = \Sigma_{12}' = \mathrm{E}(\boldsymbol{vu}') = \mathbf{S}'\mathbf{R}_{vu}$, and $\Sigma_{22} = \mathrm{E}(\boldsymbol{vv}') = \mathbf{S}'\mathbf{R}_{vv}\mathbf{S}$.

Define vectors $\boldsymbol{r} \equiv [r_1, r_2]' \equiv [z_1'\alpha_1, z_2'\alpha_2]'$ and $\boldsymbol{v} \equiv [\log(y_i) - \boldsymbol{x}'\boldsymbol{\beta_i}]$. Let $\phi(\boldsymbol{v})$ be the marginal probability density function (pdf) of $\boldsymbol{v} \sim N(0, \Sigma_{22})$ and $\phi(\boldsymbol{u}|\boldsymbol{v})$ be the conditional pdf of $\boldsymbol{u}|\boldsymbol{v} \sim N(\mu_{\boldsymbol{u}|\boldsymbol{v}}, \Sigma_{\boldsymbol{u}|\boldsymbol{v}})$, where $\mu_{\boldsymbol{u}|\boldsymbol{v}} = \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{v}$ and $\Sigma_{\boldsymbol{u}|\boldsymbol{v}} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Then, the likelihood contribution for the positive regime, where both dependent variables are positive, is given by:

(3.7) $L_1 = \phi(\boldsymbol{v}) \prod_{j=1}^{2} y_j^{-1} \int_{\boldsymbol{u}>-\boldsymbol{r}}^{+\infty} \phi(\boldsymbol{u}|\boldsymbol{v}) d\boldsymbol{u} = g(\boldsymbol{v}) \prod_{j=1}^{2} y_j^{-1} \Phi_2(\boldsymbol{r} + \mu_{\boldsymbol{u}|\boldsymbol{v}}; \Sigma_{\boldsymbol{u}|\boldsymbol{v}})$,

where $\prod_{j=1}^{2} y_j^{-1}$ is the Jacobian of the transformation from $[v_1, v_2]'$ to $[y_1, y_2]'$ and $\Phi_2(\boldsymbol{r} + \mu_{\boldsymbol{u}|\boldsymbol{v}}; \Sigma_{\boldsymbol{u}|\boldsymbol{v}})$ is the bivariate normal cumulative distribution function (cdf) with zero mean, covariance matrix $\Sigma_{\boldsymbol{u}|\boldsymbol{v}}$, and finite upper integration limits $\boldsymbol{r} + \mu_{\boldsymbol{u}|\boldsymbol{v}}$.

The second regime is one in which the values of both variables are zeros (when $z'\alpha_i + u_i \leq 0$, $i = 1, 2$). The likelihood contribution is identical to that of an all-zero regime in the bivariate probit:

(3.8) $L_2 = \int_{-\infty}^{u \leq -r} \phi(\boldsymbol{u}, \Sigma_{11}) d\boldsymbol{u} = \Phi_2(-\boldsymbol{r}; \boldsymbol{\Sigma_{11}})$,

where $\phi(\boldsymbol{u}, \Sigma_{11})$ is the marginal pdf of $\boldsymbol{u} \sim N(0, \boldsymbol{\Sigma_{11}})$. Specifically, $\phi(\boldsymbol{u}, \Sigma_{11}) = (2\pi)^{-1}|\Sigma_{11}|^{-1/2}\mathrm{e}^{-\frac{1}{2}\boldsymbol{u'}\boldsymbol{\Sigma_{11}}^{-1}\boldsymbol{u}}$.

For mixed regime, without loss of generality, denote u_i as the error term associated with the non-censored variable and u_j associated with the zero-valued variable. A mixed regime is characterized by:

(3.9) $\quad z'\alpha_i + u_i > 0 \quad \log(y_i) = x'\beta_i + v_i$

$\quad z'\alpha_j + u_j \leq 0 \quad y_j = 0.$

Let $\tilde{v} \equiv v_i$. Then $[\boldsymbol{u}', \tilde{v}]'$ is 3-variate normal with zero mean and covariance matrix $\tilde{\Sigma}$, where $\tilde{\Sigma}$ is a 3x3 sub-matrix containing the first three rows and columns of the error covariance matrix Σ in Equation (3.6). Partition $\tilde{\Sigma}$ at the third row and column such that $\tilde{\Sigma} = \begin{bmatrix} \Sigma_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{bmatrix}$.

Let $\phi(\tilde{v})$ be the marginal pdf of $\tilde{v} \sim N(0, \tilde{\Sigma}_{22})$ and $\phi(\boldsymbol{u}|\tilde{\boldsymbol{v}})$ be the conditional pdf of $\boldsymbol{u}|\tilde{\boldsymbol{v}} \sim N\,(\mu_{\boldsymbol{u}|\tilde{v}}, \Sigma_{\boldsymbol{u}|\tilde{v}})$, where $\mu_{\boldsymbol{u}|\tilde{v}} = \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\boldsymbol{v}}$ and $\Sigma_{\boldsymbol{u}|\tilde{v}} = \Sigma_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{21}$. Then the likelihood contribution for this regime is:

(3.10) $L_3 = y_i^{-1}\phi(\tilde{v}) \int_{u_i > -r_i}^{+\infty} \int_{-\infty}^{u_j \leq -r_j} \phi(u_1, u_2|\tilde{v}) du_2 du_1 = y_i^{-1}\phi(v_i)\Phi_2(\boldsymbol{D}\big(\boldsymbol{r} +$

$\mu_{\boldsymbol{u}|\tilde{v}}\big); \boldsymbol{D'\Sigma_{u|\tilde{v}}D})$,

where $\boldsymbol{D} = diag(2d_1 - 1, 2d_2 - 1),\ d_i = 1\ if\ z\alpha_i + u_i > 0$. The sample likelihood function for the censored system is the product of the likelihood contributions L_1, L_2, or L_3 across observations, depending on the regime for each observation. Maximum likelihood estimation of the model is performed in MATLAB (MATLAB, 2014).

3.2.3 Marginal Effects

Economically meaningful measures, marginal effects, are calculated based on conditional means for the joint distribution. The probability of purchase is given by:

(3.11) $\Pr(y_i > 0) = \Phi(\mathbf{z}'\boldsymbol{\alpha}_\mathbf{i})$.

Elasticity for continuous explanatory variables is defined as the change in probability of purchase, corresponding to a one-unit change in z_j. The marginal effects for indicator explanantory variables are the discrete change in purchase probabilties obtained in Equation (3.11) when the explanatory variable takes the value of one versus zero:

(3.12) $$m_i^{Prob} = \begin{cases} \frac{d\ \Pr(y_i>0)}{dz_j} = \phi(\mathbf{z}'\boldsymbol{\alpha}_\mathbf{i}) \cdot \alpha_{ij}, & if\ z_j\ continuous \\ \Phi(\mathbf{z}'\boldsymbol{\alpha}_\mathbf{i}|\mathbf{z_j}=1) - \Phi(\mathbf{z}'\boldsymbol{\alpha}_\mathbf{i}|\mathbf{z_j}=0), & if\ z_j\ binary, \end{cases}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively.

The conditional mean of expenditure y_i is (Rosiniski and Yen 2004):

(3.13) $E(y_i|y_i > 0) = \exp\left(\boldsymbol{x}'\boldsymbol{\beta_i} + \frac{\sigma_i^2}{2}\right) \cdot \Phi(\boldsymbol{z}'\boldsymbol{\alpha_i} + \rho_{ii}^{uv}\sigma_i^2)/\Phi(\boldsymbol{z}'\boldsymbol{\alpha_i})$.

Multiplying Equations (3.12) and (3.13) gets the unconditional mean of y_i:

(3.14) $E(y_i) = \exp\left(\boldsymbol{x}'\boldsymbol{\beta_i} + \frac{\sigma_i^2}{2}\right) \cdot \Phi(\boldsymbol{z}'\boldsymbol{\alpha_i} + \rho_{ii}^{vu}\sigma_i^2)$.

Let's consider a variable that enters the level equation as well as the participation equation. In this case, when deriving the semi-elasticity of conditional expected value of y_i with respect to x_j, one has to consider that vector $\mathbf{z}$ also contains x_j.

Semi-elasticity (discrete change) of the conditional mean is obtained by differentiating (differencing) Equation (3.14) with respect to variable x_j:

$$(3.15)\quad m_i^c = \begin{cases} \frac{d \ln E(y_i|y_i>0)}{dx_j} = \beta_{ij} + [\lambda(\boldsymbol{z'\alpha_i} + \rho_{ii}^{vu}\sigma_i) - \lambda(\boldsymbol{z'\alpha_i})]\alpha_{ij}\,, & if\ x_j\ continuous \\ \Delta \ln E(y_i|y_i>0) = \beta_{ij} + \Delta[\lambda(\boldsymbol{z'\alpha_i} + \rho_{ii}^{vu}\sigma_i) - \lambda(\boldsymbol{z'\alpha_i})]\,, & if\ x_j\ binary, \end{cases}$$

where α_{ij} and β_{ij} are the parameters of x_j in the participation equation and level equation for product i, respectively; $\Delta[\cdot]$indicates the difference of its argument when x_j takes the value of one versus zero. The inverse Mill's ratio is $\lambda(\boldsymbol{z'\alpha_i}) \equiv \frac{\phi(\boldsymbol{z'\alpha_i})}{\Phi(\boldsymbol{z'\alpha_i})}$.

Semi-elasticity (discrete change) of the unconditional mean with respect to x_j that enters both equations is obtained by differentiating (differencing) Equation (3.15):

$$(3.16)\quad m_i^u = \begin{cases} \frac{d \ln E(y_i)}{dx_j} = \beta_{ij} + \lambda(\boldsymbol{z'\alpha_i} + \rho_{ii}^{vu}\sigma_i)\alpha_{ij}, & if\ x_j\ is\ continuous \\ \Delta \ln E(y_i) = \beta_{ij} + \Delta[\lambda(\boldsymbol{z'\alpha_i} + \rho_{ii}^{vu}\sigma_i)], & if\ x_j\ binary. \end{cases}$$

For variables that enter the level equation only, the marginal effects for conditional and unconditional mean under both models are its parameter β_{ij} only.

Individual elasticity or discrete change is averaged over the whole sample to obtain the average marginal effect. Asymptotic standard errors for the average marginal effect estimates are obtained using the delta method (Spanos 1999).

3.3 Data and Variable Selection

The data are from the Polish household panel of about 20,000 households annually surveyed by Poland's Main Statistics Office (GUS). Despite the attempted panel structure of the survey, fewer than 36% of the households were observed for more than one year. The study uses a pooled cross-sectional sample of 77,043 observations with non-missing values for the period of 2005-2008.

The dependent variables are expenditures in the month preceding the survey on tobacco and alcohol. Positive expenditures are logarithm transformed to mitigate deviation from normality and potential heteroskedasticity.

Two variables are reported as a measure of depopulation. First, net domestic migration measures the net outflow of population from a region to other regions within Poland. Second, net international migration measures the net outflow of population from a region to other (typically EU) countries after Poland's accession to the EU in 2004.

Other demographic and socio-economic factors include: household head's gender, age, education level, marital status, and employment stability, household location, monthly income, and the number of children (age 0-18), adults (age19-60), and elders (age >60). Binary variables, the presence of children and the presence of elders, are induced from the numbers of these family members.

Table 3.1 presents a summary of sample variable statistics. Rural residents account for 37.5 % of all observed households (Village=1). Household income in the month preceding the survey averages 2,781 Polish Zloty (PLN). Nearly three out of five (59.3%) households are headed by male members, and, 67.4% of household heads are married. The proportion of household heads with secondary or higher education is 40.7%. The average household head's age is 51.1 years. In term of employment stability, 26.6% of household heads are permanently employed or contract employees. The average household size is 2.98 family members, with the average number of children (age 0-18 years), adults (age 19-60 years), and elders (above 60 years of age) 0.72, 1.80, and 0.45 per household, respectively. Households with the presence of children and elders account for 42.1% and 33.7%, respectively. On average, net migration inflow from one Polish region to another domestic region averages 1,352 persons over all 16 administrative regions of Poland.

Outflow from a Polish region to a foreign country averages 1,565 persons annually. The proportions of households observed in each year are fairly balanced, with 25.1% in 2005 and 2006, and 24.9% in 2007 and 2008.

The percent of households who bought tobacco and alcohol in the month preceding the survey is 36.3% and 56.2%, respectively. This paper loosely refers to positive expenditures on tobacco and alcohol as smoking and drinking, respectively. Average spending on tobacco and alcohol are PLN41.0 and PLN28.5, respectively. Conditional on purchase, households on average spend PLN112.75 per month on tobacco and PLN50.63 per month on alcohol (figures not reported in Table 3.1). Lastly, household observations distribute quite evenly over the four-year period, with about one quarter of households observed in each year. Also, the household characteristics are also very similar over time (statistics not shown in Table 3.1).

3.4 Results

For the censored system (Table 3.2), parameter estimates are obtained by maximum likelihood estimation. The left panel of Table 3.2 reports parameter estimates for the participation decisions and the right panel reports on the level decisions. Additionally, Table 3.2 reports estimated error correlation coefficients among all equations.

The correlation coefficients for the participation and level equations are estimated to be positive and statistically significant at the 5% level. The correlation coefficient between the decisions to smoke and drink is estimated as high as 0.225 with a *p*-value lower than 1%. The error correlation between the expenditures on tobacco and alcohol is 0.086 (*p*-value <1%). This positive correlation between the behaviors of smoking and drinking are probably due to their addictive nature. The non-zero correlation between the decision to smoke (drink) and the expenditure on

tobacco (alcohol) indicates the presence of sample selection. The correlation coefficients across the decision to smoke and the expenditure on alcohol as well as the decision to drink and the expenditure on tobacco are statistically different from zero as well. These results validate the necessity of a system approach.

The parameter coefficient estimates and their corresponding standard errors show whether a variable plays a statistically significant role in each equation. However the coefficient estimates *per se* do not directly quantify the magnitude of each variable's effect. Therefore, this study also computed a more meaningful measure, marginal effect. The marginal effects on the probability of purchase and on the expenditure level are reported in Tables 3.3 and 3.4, respectively. Furthermore, Table 3.4 reports marginal effects on expenditure amount based on both the conditional and conditional means. Marginal effects based on the conditional mean provide insight about the population who smoke/drink, while the latter facilitate correct inference about the general population. One should adopt the latter if a policy targets to reduce tobacco and alcohol use by the general population, instead of singling out drinkers/smokers (which is much harder). In this study, these two sets of marginal effects have the same signs, and approximate each other in values and significance level. However, it is still important to use the correct measure. Small difference in the absolute values could mean drastic difference economically, especially when it is about a large population and a large spending on these two products.

An examination of the marginal effects of each variable provides additional insights. For the expenditure level, this study only illustrates through the marginal effects based on conditional means, since the two sets of marginal effects have similar patterns.

An increase of 1000 PLN in income is associated with a 5.1% higher probability of alcohol purchase. But the effect on the decision to smoke has not been confirmed. A decision to smoke

may be made early in life when income as compared to tobacco price is not a major constraint, especially because cigarettes are offered in a wide price range in Poland. Once the smoking habit has established, the higher expenditure on alcohol in households with higher incomes reflects the ability to purchase better quality and quantity.

Residency in rural areas is associated with a higher propensity of 1.89% and 11.22% for drinking and smoking, respectively. However, rural residents' expenditure on tobacco and alcohol are lower than those of their urban counterparts. These result reflect not only similar preferences but also likely differences in lifestyle between rural and urban residents. An earlier study found differences in alcohol and tobacco demand between rural and urban residents (Florkowski and McNamara 1992).

The gender of household heads does not have a statistically significant effect on the decision either to drink or smoke. The expenditures of these households on either product are slightly lower than those headed by males. But the differences are negligible small (less than 0.1%). The marital status of household heads do not affect the decisions of smoking and drinking, either. But the expenditure on both products are slightly higher for those with married household heads, indicating a possibly different lifestyle. Higher household head's education level decreases the probability of tobacco purchase. This result is consistent with the significance of education in forming healthy consumption choices. However, more formal education is associated with slightly higher (5.1%) probability of alcohol purchase. Conditional on purchase, household heads with higher education levels on average spend a little more on both products.

Families with older household heads are more likely to smoke, but less likely to drink. The result coincides with decreasing smoking rates among young consumers as compared to their parents or grandparents. The expenditures of households headed by older consumers are higher on

tobacco but lower on alcohol. Household head's status of employment does not play a statistically significant role on either product. But household heads with more stable employment spend more on both products. This is interesting because some earlier studies reported mixed effects of employment on alcohol and tobacco use (e.g. Bilgic and Yen, 2015).

Larger households are more likely to buy tobacco but less likely to buy alcohol. A larger family in Poland may consist of multiple generations and more adults, who learn the same habit of smoking, sometimes by easy access to cigarettes. Once a household decides to buy alcohol, they spend more than a smaller household. This is not surprising, since larger household size usually means more drinkers.

The presence of elders is associated with a higher probability of smoking. Smoking peaked in Poland a couple of decades ago and older individuals picked up the smoking habit much earlier. However, the presence of the elderly lowered the propensity for drinking. This pattern has similarity to influence of the household head's age on the decision to purchase tobacco and alcohol.

A household with children is less likely to buy alcohol, but interestingly, they are associated with higher probabilities of tobacco purchase. This result seems to contradict the food demand literature where households with children usually make healthier food choices. But the consumption of tobacco and alcohol is different from typical food consumption. People might form the habit of drinking or smoking before they have children and this habit persists because of its addictive nature.

Outmigration, a special issue in Poland, has intensified as a result of the domestic economy adopting its structure to market-driven resource allocation and outside opportunities resulting from the EU accession in 2004. Outmigration, measured by domestic and international net outflow, has somewhat mixed effects. Both domestic and international outmigration is associated with lower

probability of smoking. Migrating workers are usually young, better educated in seek of employment with higher wages. They generally migrate to regions or countries that are economically better developed than Poland. Higher educational attainment and younger age are both associated with a lower probability of smoking, while migrants to better developed areas in Poland or the EU quite possibly become exposed to different lifestyle and cultural values. These values are communicated back to their families staying behind and affect the decision to smoke. Similarly, both migration measures are associated with lower tobacco expenditure. In contrast, households in regions with higher outmigration, either domestically or internationally, are more likely to drink and also on average spend more on alcohol purchase. A large number of migrants to EU countries end up in countries where people traditionally have much higher beer (for example, Germany, Ireland, the Netherlands) or wine (for example, France, Italy or Spain) consumption. Migrants likely absorb local lifestyle and consumption habits including drinking, which they project to families left behind.

Lastly, the effect of year indicator is quite large. The year indicator does not only capture the changes in consumption over time, but also captures the effect of influential factors that were not explicitly observed and modelled in the empirical analysis, for example, the changes in Poland's GDP and, thus, purchasing power. The growth rates of Poland's GDP were 5.1%, 3.5%, 6.2%, 7.2%, and 3.9% in 2004 to 2008, respectively. There was a slowdown in GDP growth between 2004 and 2005, then acceleration in 2006 and 2007, and lastly a sharp slowdown in 2008. This pattern is partially reflected in the effect of year indicators. Compared to 2005, the years of 2006 and 2007 are associated with higher probability of purchase as well as higher expenditure, while the year of 2008 witnessed a significant slowdown in the case of tobacco consumption and a sharp decrease in the case of alcohol consumption.

3.5 Conclusions and Discussions

Harmful use of tobacco and alcohol is one of the main risk factors for adverse health and social consequences. With the growing attention to tobacco- and alcohol-related social and health problems and public policy campaigns against tobacco and alcohol use, an analysis of the determinants of household tobacco and alcohol consumption remains important. This study takes advantage of household survey data collected by Poland's Main Statistical Office that is not publicly available. The empirical investigation applies a censored system. This multivariate sample selection model addresses the censoring feature of the survey data. It also allows error correlation among all equations to consider sample selection and the possible correlation between tobacco and alcohol use due to their addictive nature.

The empirical model uses three categories of explanatory variables. Household features include household income, location (rural vs. urban residence), and household size and structure. Household head characteristics are age, gender, education level, marital status, and employment status. Lastly, worker outmigration (both domestic and foreign), a special issue in Poland, is investigated.

The empirical estimation indicates that the decisions to smoke and drink and their expenditure levels are indeed positively correlated. In the case of tobacco use, rural residence, older household heads, larger household, and the presence of elders are associated with a higher probability of tobacco purchase. Married household heads, higher education level, and higher outmigration, both domestically and internationally, are less likely to be associated with buying tobacco. The presence of children, unfortunately, does not play a role in reducing the likelihood of tobacco use. This potentially exposes children to second-hand tobacco smoke.

In the case of alcohol purchase, rural residence, higher income, higher education level, and higher domestic outmigration positively affects the likelihood of drinking. Meanwhile, larger households as well as those with children and elders are less likely to buy alcohol.

Some explanatory variables have similar effects on the expenditure on tobacco and alcohol. Rural residence and male household heads are associated with lower expenditure on both products. Married household heads and those with higher education, on the contrary, spend more on both products. The rest of the explanatory variables have mixed effect across tobacco and alcohol expenditure. Higher income increases the expenditure on alcohol, but it does not affect tobacco use, a habit probably established earlier in life regardless of income level. Older household heads spend more on tobacco but less on alcohol. This is consistent with the observed trends of lifestyle shifts in Poland. Higher outmigration is associated with lower expenditure on tobacco, but higher spending on alcohol. This might reflect possible changes in lifestyle and, thus, changes in consumption patterns due to the exposure to different lifestyles and culture. Overall, this study's findings reveal determinants of household consumption of tobacco and alcohol, and identify household features that are likely to be associated with large expenditures. Such households or individuals from such households are a potential target for policies aimed at reducing the harmful effects of alcohol and tobacco consumption.

Table 3.1 Summary of Descriptive Statistics of Sample Variables

Variable	Description/Unit	Mean	Std Dev
	Dependent Variables		
Smoke	1, if a household buys tobacco, 0 otherwise	0.363	-
Drink	1, if a household buys alcohol, 0 otherwise	0.562	-
Tobacco	Expenditure on tobacco in the month preceding survey, in PLN	40.984	78.917
Alcohol	Expenditure on alcohol in the month preceding survey, in PLN	28.470	60.472
	Demographic, Socio-Economic Factors / Explanatory Variables		
Village	1, if a household residents in village, 0 otherwise	0.375	-
Income	Household income in the month preceding survey, in 1000 Polish Zloty (PLN)	2.781	2.205
Male	1, if the household head is male, 0 otherwise	0.593	-
Married	if the household head is married, 0 otherwise	0.674	-
HighEduc	1, if the household head has secondary or higher education, 0 otherwise	0.407	-
Age	Household head's age, in years	51.146	15.210
Employed	1 if household head is permanently employed or contract employee, 0 otherwise	0.266	-
Hhsize	Number of family members in a household	2.981	1.531
NKid	Number of children (under 18)	0.723	1.040
N1960	Number of adults 60 or under 60 years old	1.804	1.191
N60above	Number of elders above 60	0.453	0.696
DKid	1 if children are present in a household, 0 otherwise	0.421	-
DElder	1 if elders (above 60) are present in a household, 0 otherwise	0.337	-
OUTD	Net migration domestically to other regions in Poland, in 1000	-1.352	5.714
OUTF	Net migration international to other countries, in 1000	1.565	2.108
YR05	Baseline, 1 if observed in 2005, 0 otherwise	0.251	-
YR06	1 if observed in 2006, 0 otherwise	0.251	-
YR07	1 if observed in 2007, 0 otherwise	0.249	-
YR08	1 if observed in 2008, 0 otherwise	0.249	-

Note: N=77,043

Table 3.2 Maximum-likelihood Estimates for Censored System of Tobacco and Alcohol Consumption

	Binary Decision Of Smoking	Binary Decision of Drinking		Expenditure on Tobacco	Expenditure on Alcohol
	Coeff. (Std. Err.)	Coeff. (Std. Err.)		Coeff. (Std. Err.)	Coeff. (Std. Err.)
Constant	-0.563 (0.015)**	-0.379 (0.022)**	Constant	5.779 (0.038)**	4.230 (0.041)**
Village	0.053 (0.010)**	0.298 (0.010)**	Village	-0.029 (0.018)*	-0.018 (0.015)
Income	-0.003 (0.003)	0.138 (0.012)**	Income	0.049 (0.007)**	0.051 (0.008)**
Male	0.0003 (0.001)	0.003 (0.001)**	Male	-0.006 (0.002)**	-0.005 (0.001)**
Married	-0.007 (0.003)**	-0.006 (0.003)**	Married	0.003 (0.004)	0.013 (0.003)**
HighEduc	-0.075 (0.012)**	0.135 (0.014)**	HighEduc	0.174 (0.021)**	-0.111 (0.018)**
Age	0.218 (0.011)**	-0.035 (0.011)**	Age	-0.207 (0.018)**	-0.108 (0.015)**
Employed	-0.004 (0.013)	0.009 (0.014)	Employed	0.021 (0.023)	-0.005 (0.019)
Hhsize	0.472 (0.004)**	-0.789 (0.052)**	N1960	-0.009 (0.011)	0.089 (0.012)**
DKid	0.405 (0.012)**	-0.811 (0.055)**	N60above	-0.031 (0.012)**	0.112 (0.013)**
DElder	0.384 (0.005)**	-0.874 (0.052)**			
OutD	-0.034 (0.013)**	0.021 (0.013)*	OutD	0.119 (0.022)**	0.044 (0.017)**
OutF	-0.056 (0.013)**	0.007 (0.014)	OutF	0.165 (0.023)**	0.048 (0.018)**
YR06	-0.141 (0.014)**	0.045 (0.016)**	YR06	0.147 (0.022)**	-0.131 (0.018)**
YR07	-0.211 (0.018)**	-0.216 (0.022)**	YR07	0.329 (0.020)**	0.119 (0.017)**
YR08	-0.328 (0.001)**	0.801 (0.052)**	YR08	-0.107 (0.008)**	0.003 (0.007)

Correlation Coefficient Estimates			
	Coeff. (Std. Err.)		Coeff. (Std. Err.)
Rho.Smoke.Drink	0.225(0.009)**	Rho.Smoke.Tobacco	-0.951(0.132)**
Rho.Tobacco.Alcohol	0.086(0.009)**	Rho.Smoke.Alcohol	-0.033(0.009)**
Sigma.Tobacco	1.605(0.010)**	Rho.Drink.Tobacco	-0.143(0.011)**
Sigma.Alcohol	1.408(0.012)**	Rho.Drink.Alcohol	-0.833(0.091)**

** Significant at 5%.
* Significant at 10%.

Table 3.3 Marginal Effects on Purchase Likelihood

	Binary Decision of Smoking	Binary Decision of Drinking
	Coeff. (Std. Err.) [1]	Coeff. (Std. Err.)
Constant	-20.568 (0.468)**	-13.999 (0.874)**
Village	1.891 (0.398)**	11.221 (0.461)**
Income	-0.081 (0.444)	5.102 (1.050)**
Male	0.010 (0.223)	0.093 (0.284)
Married	-0.232 (0.280)	-0.207 (0.446)
HighEduc	-2.626 (0.443)**	5.083 (0.634)**
Age	7.954 (0.408)**	-1.287 (0.422)**
Employed	-0.156 (0.495)	0.325 (0.505)
Hhsize	17.083 (0.297)**	-29.210 (2.550)**
DKid	13.953 (0.467)**	-27.709 (1.387)**
DElder	13.228 (0.207)**	-24.604 (1.480)**
OutD	-1.202 (0.473)**	0.761 (0.477)*
OutF	-1.950 (0.479)**	0.252 (0.511)
YR06	-5.228 (0.485)**	1.631 (0.649)**
YR07	-7.687 (0.629)**	-8.117 (0.844)**
YR08	-10.261(0.404)**	11.501 (0.929)**

[1] The coefficient estimates are in percentage and standard errors were multiplied by 100
** Significant at 5% level
* Significant at 10% level

Table 3.4 Marginal Effects on Expenditure Level

Semi-elasticity of Conditional Mean			Semi-elasticity of Unconditional Mean		
	Tobacco Expenditure	Alcohol Expenditure		Tobacco Expenditure	Alcohol Expenditure
	Coeff. (Std. Err.)	Coeff. (Std. Err.)		Coeff. (Std. Err.)	Coeff. (Std. Err.)
Constant	-12.956 (0.535)**	2.215 (0.103)**	Constant	-12.364 (0.520)**	2.484 (0.089)**
Village	-0.087 (0.010)**	-0.323 (0.010)**	Village	-0.050 (0.003)**	-0.144 (0.004)**
Income	-0.024 (0.088)	0.787 (0.052)**	Income	-0.022 (0.085)	0.689 (0.044)**
Male	-0.006 (0.001)**	-0.009 (0.001)**	Male	-0.006 (0.0003)**	-0.007 (0.0004)**
Married	0.011 (0.003)**	0.010 (0.003)**	Married	0.006 (0.001)**	0.006 (0.001)**
HighEduc	0.252 (0.012)**	0.049 (0.014)**	HighEduc	0.202 (0.004)**	0.130 (0.005)**
Age	6.889 (0.344)**	-0.293 (0.054)**	Age	6.665 (0.333)**	-0.268 (0.046)**
Employed	0.025 (0.013)**	0.012 (0.013)	Employed	0.022 (0.004)**	0.017 (0.005)**
N1960	-0.010 (0.011)	0.089 (0.012)**	N1960	-0.010 (0.011)	0.089 (0.012)**
N60above	-0.029 (0.012)**	0.111 (0.013)**	N60above	-0.029 (0.012)**	0.111 (0.013)**
OutD	-0.978 (0.408)**	0.154 (0.061)**	OutD	-0.944 (0.395)**	0.139 (0.052)**
OutF	-1.611 (0.413)**	0.085 (0.066)*	OutF	-1.555 (0.399)**	0.080 (0.056)*
YR06	0.314 (0.014)**	0.126 (0.016)**	YR06	0.212 (0.004)**	0.152 (0.006)**
YR07	0.574 (0.018)**	0.572 (0.022)**	YR07	0.423 (0.005)**	0.442 (0.008)**
YR08	0.210 (0.001)**	-0.904 (0.054)**	YR08	0.041 (0.001)**	-0.226 (0.002)**

** Significant at 5% level

* Significant at 10% leve

CHAPTER 4

A HOLISTIC PICTURE OF HOUSEHOLD AT-HOME FRESH PRODUCE CONSUMPTION

This study provides a holistic profile of fresh produce choices and expenditures, including expenditure on fresh produce, frequency of purchase, variety of selection, and use of deals and coupons. The presentation provides a holistic picture of consumer disadvantage in terms of fresh produce consumption and take an all-inclusive approach so as to seek out commodities as well differences in fresh produce shopping behaviors across four consumer groups. A profile of consumers by consumer group was developed using Nielsen Homescan panel in 2014. Socio-economic and demographic features indicating consumer disadvantage were given special attention. Possible empirical analyses are discussed for future research.

4.1 Introduction

The rising rate of obesity in the United States has led to increasing attention to the consumption of fruits and vegetables. The consumption of fruits and vegetables can play an important part in reducing incidence of overweight and obesity, as they tend to have fewer calories per serving than other foods. Encouraging Americans to eat more fruits and vegetables has been a central theme of Federal dietary guidance for more than a decade. At the same time, the growth of international trade and the retail sector offers increasing quantity and variety of fresh produce (Guthrie et al. 2005). Yet, despite the health benefit, favorable market trends, as well as consistent Federal

recommendations, most American fall short of their recommended fruit and vegetable intake. USDA food supply data indicate that Americans consume 1.4 servings of fruit daily, less than half the 4 servings or 2 cups recommended in the 2005 *Dietary Guidelines* for adults eating 2,000 calories per day. Vegetable consumption is 3.7 servings per day, also below the recommended 5 servings or 2.5 cups per day (ERS, USDA, 2016).

Disadvantaged groups are least likely to have fruit and vegetable intakes that are consistent with healthy eating messages (Giskes et al. 2002; Mishra et al., 2002; Turrell et al., 2002). Many studies have set out to investigate food shopping patterns of one specific group of disadvantaged consumers, for example, the low-income (e.g. Clifton 2004), the elderly (e.g. Wilson, Alexander and Lumbers 2004), women (e.g. Herman et al., 2008), and ethnic minorities (e.g. Zenk 2005). However, disadvantaged consumers are not well defined. And studies that take a more holistic view have been lacking. This study intends to present a holistic picture of consumer disadvantage and take an all-inclusive approach so as to seek out commonalities as well as differences in fresh produce shopping behavior across different characteristics of disadvantaged consumers. Such findings provide key insights to developing a solution to the rising rate of obesity in the United States.

4.2 What Are the Characteristics of Disadvantaged Consumers?

A review of the marketing and retail literatures reveals a weakly developed and fragmented picture of consumer disadvantage (Woodliffe 2004). The conceptualizing and theorizing of consumer disadvantage is deterred perhaps by its complex and multi-dimensional nature (Bromley and Thomas 1993; and Woodliffe 2004). While this study does not attempt this task, a review of the

characteristics of disadvantaged consumer remains important before one investigates their food purchase behaviors.

It is generally accepted within the existing, albeit sparse, literature that the notion of consumer disadvantage hinges on inequality. Certain consumers are unequal in the marketplace as they belong to socially disadvantaged groups such as the elderly, the disabled or low income households.

Fair Trade Commission (2016) exemplifies that "some consumers may be disadvantaged or vulnerable in some marketplace situations if they: Have a low income; Are from a non-English speaking background; Have a disability …; Have a serious or chronic illness; Have poor reading, writing and numerical skills; Are homeless; Are very young; Are old". Past studies on this topic have tended to focus on socially disadvantaged individuals and label them disadvantaged consumers before assessing their shopping experience (Woodliffe 2004). However, not all consumers with these characteristics face additional constraints in food purchase. Bromley and Thomas (1993) provide a useful point to disentangle and structure the various facets of consumer disadvantage. These authors consider consumer disadvantage to emerge from two interrelated dimensions, social disadvantage and poor personal mobility. Taking their ideas a step further, the phenomenon can be viewed as consisting of potential causes and manifestations (Woodliffe 2004). Potential causes of consumer disadvantage are linked to membership of social disadvantage groups, and help to identify who is likely to be disadvantaged as a consumer. Manifestations are concerned with the forms that disadvantage may take in a retail context, for example, paying higher prices, lack of variety and selection and poorer quality (Market Behavior Ltd, 1983).

Following Bromley and Thomas' (1993) idea, this study considers the potential cause and manifestation of consumer disadvantage at the same time. Specifically, this study first examines

households' consumption of fresh fruits and vegetables against household socio-economic and demographic features, focusing on the commonalities and differences in shopping behaviors associated with the characteristics of disadvantaged consumers. To take the analysis a step further, this study also assesses consumers' shopping experience of fresh produce in terms of shopping frequency, variety of selection, use of coupons and deals, and level of expenditures. Definition of these measurements will be discussed in the next session.

The membership of social disadvantaged groups provides a convenient and widely available indicator for possible disadvantaged consumers. Socio-economic and demographic information available in Nielsen Homescan panel for identifying possible disadvantaged consumers are: low income, less education, ethnic minorities, age, gender and employment status of household heads, who are usually the primary food buyers.

In the analysis of this study, household income was categorized into four levels using three cutoff values: one, two, and four times of the 2012 Federal Poverty Line (FPL) (HHS, 2012). The 2012 Federal Poverty Line (FPL) was used because Nielsen Homescan panelists were asked to report their combined total household annual income of two years prior to the panel year. Twice of the FPL was used as a cutoff, because research shows that, on average, families need an income of about twice the poverty line to cover basic expenses. Households with income below four times of the FPL may be eligible for tax credits and/or other types of government assistance. *The elderly* are those age 65 years or older. *The less educated* are those with some high school education or high school diploma. Three groups of ethnic minorities, *African American*, *Asian*, and *other* are also investigated.

Lastly, it is worthy to emphasize that these characteristics of consumer disadvantage are not exhaustive or exclusive. One socially disadvantaged feature, for example, low income, is very

likely to be correlated with another, for example, low education. Therefore, this study also divides fresh produce buyers into four groups and develops a profile of consumers based on their socio-economic and demographic features, focusing on those closely related to consumer disadvantage as defined previously.

4.3 The Sample Data and Description

This study analyzes the 2014 Nielsen Homescan panel data. According the Nielsen, the panel consists of representative U.S. households that provide food purchase data for at-home consumption. In general, panelists report their purchases weekly by scanning either the Universal Product Code (UPC) or a designated code for random-weight products of all their purchases from retail outlets.

This study analyzes purchases belonging to the category of fresh produce reported by 45,328 panelists (households) in 2014. The Homescan data include product characteristics and promotion information, as well as detailed socio-economic and demographic information for each household.

In the analysis of this study, consumer shopping experiences were profiled by four measures on per household basis. The *per capita expenditures* were per capita actual spending after adjusting for the value of coupons and the effect of household size.

The *frequency of purchase* was calculated as number of shopping trips for fresh produce. More frequent purchase of fresh produce usually means larger consumption per day and a more even distribution of consumption over time. In addition, it might also somewhat indicate the freshness and quality of fresh produce.

Another dimension of fresh produce consumption is *variety of selection*. Different foods contain different nutrients and other substances. The Food Guide Pyramid recommends to eat a variety of fruits and vegetables daily to ensure the intake of different nutrients. The *variety of selection* in this study was calculated as the number of distinct product (UPC)[1] bought by a household during the year.

Cost was the most commonly and extensively described barrier to purchasing fresh fruits and vegetables by the low income population (e.g. Haynes-Maslow et al. 2011). The use of coupons and price discounts not only helps to lower the cost, but also shifts food choices. Price discounts and coupons are proposed as method to promote the consumption of fresh produce in government cost assistant program (e.g. The Massachusetts Farmers Market Coupon Program distributes coupons for fresh fruits and vegetables redeemable at Farmer's Markets). This study examines the use of deals and coupons by different consumer groups. The findings might help to distinguish target groups. The *use of deals and coupons* was expressed as proportion of households that used deals or coupons. Another measure of the use of deals was calculated as proportion of purchase records considered as deals on per household basis.

4.4 What Factors Affect Fresh Produce Purchase?

Of all demographic features, household income and race seem to be most correlated with fresh produce consumption. Per capital spending is positively correlated with household income level (Figure 4.1). Households that are under FPL spent only $75.83 on fresh produce on per capita

[1] Different UPC does not necessarily mean different types of fresh produce. It might be the same type of fresh produce by different brands and/or in different package.

basis in 2014, while those above four times of the FPL spent 170% of this amount, reaching $130.20 per person.

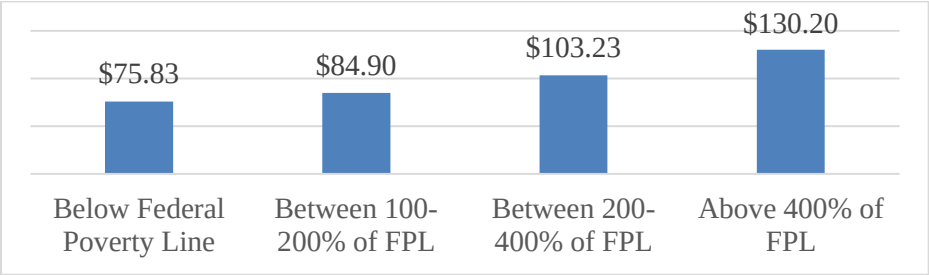

Figure 4.1 Average per capita spending on fresh vegetables for home consumption by income class[2]

The consumption of fresh produce also differs significantly by ethnic group (Figure 4.2). Asian Americans spent the most food dollars ($115.35) to purchase fresh produce on a per capita basis. African American consumers on average spent significantly less, about 80% of the per capita spending by Asian Americans.

[2] Group means are pairwise significantly different from each other by pairwise t-test upon logarithm transformation of per capita spending. The same method was used for tests hereafter unless otherwise indicated.

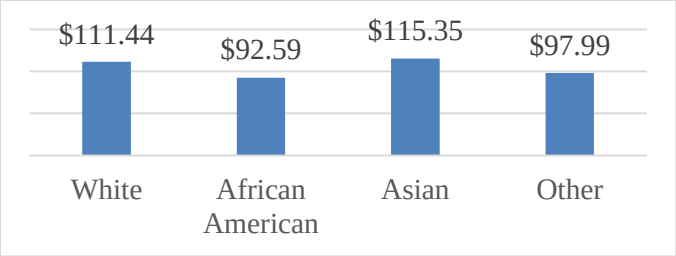

Figure 4.2 Average per capita spending on fresh produce for home consumption by race

It is interesting to note that average per capita spending on fresh produce for home consumption exhibited a U-shape relationship with the age of household heads (Figure 4.3). Household heads under 30 years old spent a moderate amount on fresh produce ($89.06 per person). The per capita spending reaches the lowest for the age group of 30-39 years ($78.81), and then climbed up for the older household heads, reaching $115.51 and $127.83 for age groups of 50-64 years and 65 + years, respectively.

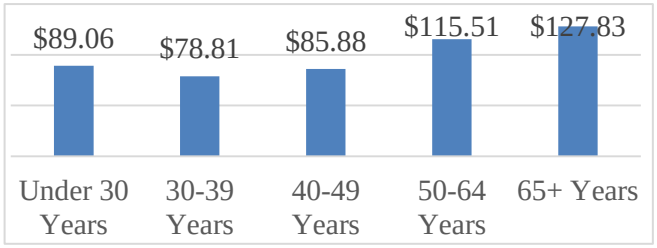

Figure 4.3 Average per capita spending on fresh produce for home consumption by age of household head[3]

[3] For married-couple households, demographic features of female head were used if not otherwise indicated.

The consumption of fresh produce varies by region (Figure 4.4). Results show that households in the Midwest and South spent less than those in the Northeast and West.

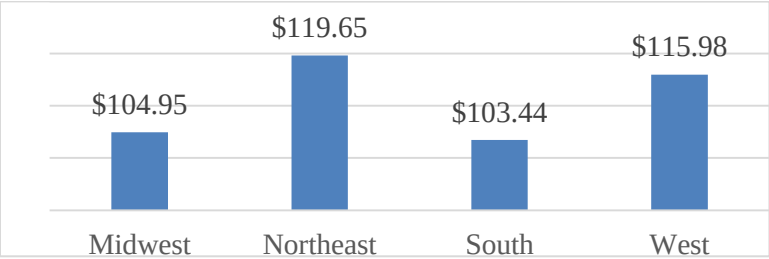

Figure 4.4 Average per capita spending on fresh produce for home consumption by region

4.5 A Profile of Consumers by Consumer Group

The analysis in this study categorizes each household into consumer or non-consumer group according to whether or not the household purchased fresh produce. In the sample data, 98.2% households purchased fresh produce. The consumer households are classified into one of three groups based on sample distribution of per capita spending. Consumer households that spent less than the first quartile ($31.84) were defined as non-frequent buyers. Households, which spent between the first and third quartile ($31.84 - $142.09), were defined as frequent buyers. Finally, those which spent in excess of $142.09 on per capita basis were defined as very frequent buyers.

As shown in Table 4.1, the frequency of purchase differs significantly across consumer group. The average number of trips per year by non-frequent, frequent, and very-frequent buyers were 11.4, 33.5, and 61.2, respectively. That is, non-frequent buyers, on average, made 0.95 trip to buy fresh produce per month. Frequent buyers shopped for fresh produce about every one and

a half weeks, while very-frequent buyers, on average, bought fresh produce about 1.2 times per week. A first glance seems to suggest that non-consumers and non-frequent buyers might be risk of low consumption.

Not surprisingly, the varieties of selection are positively correlated with the level of spending. Non-frequent buyers, on average, purchased 4.8 and 7.6 distinct fresh fruit and vegetable products, respectively. The numbers of distinct fresh fruits purchased by frequent and very-frequent buyers are 13 and 23.2, respectively. That is, 2.7 and 4.8 times of the varieties of selected fruits by non-frequent buyers. And the numbers of distinct fresh vegetable products are 7.6, 19.3, and 32.2 for non-frequent, frequent and very-frequent buyers, respectively.

There are huge gaps in average per capita spending across consumer group. Non-frequent buyers on average spent $8.92 on fresh produce, while frequent and very-frequent buyers spent as much as 8.5 and 30 times, respectively, reaching $76.05 and $267.96 per family member.

The proportion of households that used deals and coupons climbs up from non-frequent to very-frequent buyers. Interestingly, the proportion of purchases considered as deals is the highest for non-frequent buyers, followed by frequent and very-frequent buyers in that order. This result indicates that price discounts are more likely to shift non-frequent buyers' food choice decisions.

Comparing demographic information across consumer groups offers further insights in terms of how fresh produce consumption is related to these characteristics. Table 4.2 summarizes selected household characteristics by consumer group. Non-consumers and non-frequent buyers comprise relatively larger proportions of low-income households than frequent buyers. However, there is still a considerable proportion of high income households in the non-consumer groups. Higher income consumers might dine out more often, and thus purchase less fresh produce for at-home consumption. But it is important to note that away-from-home foods might be associated

with risks of poor diet quality. For example, Lin, Guthrie and Blaylock (1996) showed that away-from-home foods consumed by children were higher in total fat, saturated fat and sodium and lower in fiber and calcium than home foods. Additional, the analysis did not include fresh produce from home and community gardens or processed vegetables and fruits, such as frozen, canned, and dried produce.

Frequent and very-frequent buyers comprise a large proportion of those who have at least college degree, while a large proportion of non-frequent buyers and non-consumers either have some high school or a high school diploma. Different consumer groups seem to have similar regional distribution. With respect to the age of household head, very-frequent buyers comprise the least proportion of the youngest households (household head age <30 years), and the largest proportion of the elderly (household head age ≥ 65 years), compared to other consumer groups. Frequent and very-frequent buyers have the largest proportion of whites relative to other consumer groups, while a larger proportion of African American consumers belong to the non-consumer and non-frequent buyer groups. Interestingly, non-frequent buyers have the largest proportion of households with children, followed by frequent buyers, non-consumers, and lastly, very-frequent buyers.

4.6 Summary and Discussion

This study used the 2014 Nielsen Homescan panel to analyze consumer purchase patterns of fresh produce. In additional to expenditures, frequency of purchase, variety of selection, and use of deals and coupons are also investigated to provide a holistic picture of consumer purchase patterns.

Our analysis shows that per capita spending on fresh produce varies drastically by consumer group. The frequent and very-frequent buyers spent as much as 8.5 and 30 times of the

amount by non-frequent buyers. The frequency of purchase is once per month, once per 1.5 weeks, and 1.2 times per week for non-frequent, frequent, and very-frequent buyers, respectively. Heavier users also have a wider variety of selection in terms of number of distinct products as well as a larger proportion of households who used deals and coupons. Interestingly, out of their purchases, non-frequent buyers bought the largest proportion of products with price discounts, followed by frequent buyers, and lastly, very-frequent buyers. This result indicates that households with less consumption of fresh produce are proportionally more likely to use price discounts. This finding is favorable to the proposal of promoting the consumption of fresh produce through price discounts and coupons. It is recognized that the price discounts in the sample data are in-store commercial promotions and, thus, differ from those proposed in literature for government assistance program. Studies specifically designed for the nature of government distributed coupons should be conducted before one can make conclusive inferences.

A profile of consumers by consumer group was also developed. This study shows that the less educated, the low-income, African American consumers, and the youngest households (under 30 years) are at higher risk of low at-home consumption. Very-frequent buyers have the largest proportion of the elderly, followed by non-consumers, frequent buyers and non-frequent buyers. The elderly do not seem to be at risk of low consumption relative to the youngest households.

The current analysis was conducted on an aggregate level. The Nielsen Homescan data could facilitate analyses in more details. For example, the consumption of fresh produce could be broken down into 10 types of fresh fruits and 16 types of vegetables. Given the wide selection of products, it is expected that a large proportion of households has zero spending on each product. A censored demand system therefore is appropriate for accommodating censoring and multiple-products purchase decisions. The application of an empirical model also quantifies the effect of

each socio-economic and demographic characteristic, providing further insights for understanding household consumer behaviors. This is an interesting potential area for future research.

Lastly, the analysis in this study attempts to reveal household characteristics associated with low at-home consumption of fresh produce. The consumption of substitute products, such as away-from-home food, canned, frozen and dried produce, and fresh produce from gardens, were not investigated. While the Nielsen Homescan data does not facilitate the analysis of fresh produce consumption away-from-home or those from community gardens, the data does include purchase records of frozen, canned and dried fresh produce. Future research in this direction will provide a holistic picture of vegetable and fruit consumption in various forms.

Table 4.1 A Profile of Fresh Produce Purchase by Consumer Group

Category	Non-consumers	Non-frequent buyers	Frequent buyers	Very-frequent buyers	Total
Average Frequency of Purchase [1]		11.40	33.50	61.20	34.90
Average Variety of selection [2]					
Fresh fruits		4.80	13.00	23.20	13.50
Fresh vegetables		7.60	19.30	32.20	19.60
Average Per Capita Spending ($)		8.92	76.05	267.96	107.25
Use of Deals and Coupons					
% of households that used deals		55.25	75.82	82.73	71.10
% of purchases considered as deals		37.47	27.80	22.70	28.20
% of households that used coupons		15.32	32.83	46.96	31.42
Number of Households	**825**	**11125**	**22252**	**11126**	**45328**

Note: buyer groups are classified based on the first and third quartile of nonzero average per capita spending on fresh produce. Q1=$31.835 and Q3=$142.085.
[1]: Measured as number of shopping trips
[2]: Measured as number of distinct products (UPC) purchased

Table 4.2 Selected Household Characteristics by Consumer Group

Category	Non-consumers	Non-frequent buyers	Frequent buyers	Very-frequent buyers	Total
Income Class (%)					
Below FPL	11.88	10.26	5.67	3.60	6.40
Between 100-200% of FPL	24.85	24.18	18.89	12.38	18.70
Between 200-400% of FPL	28.97	34.52	35.46	31.83	34.22
Above 400% of FPL	34.30	31.05	39.98	52.19	40.68
Education Level (%)					
High school diploma or less	29.82	30.71	28.39	23.18	27.70
Some college	28.61	30.62	30.05	29.00	29.90
College degree	29.33	28.06	29.57	31.37	29.64
Post college degree	12.24	10.60	12.00	16.46	12.75
Region (%)					
Midwest	24.36	25.19	27.06	24.57	25.94
Northeast	18.42	15.60	16.67	19.18	17.06
South	36.61	39.76	36.88	34.49	36.99
West	20.61	19.46	19.39	21.76	20.01
Age of Household Head (%)[1]					
Under 30 Years	2.55	2.93	2.40	1.67	2.36
30-39 Years	9.33	14.11	9.92	6.33	10.06
40-49 Years	18.79	23.12	18.18	12.02	17.89
50-64 Years	42.06	40.38	43.56	47.35	43.68
65+ Years	27.27	19.46	25.94	32.64	26.02
Race					
White	76.48	78.96	83.58	84.46	82.53
African American	14.67	13.65	9.57	7.03	10.04
Asian	3.15	2.60	2.77	4.06	3.05
Other	5.70	4.80	4.08	4.45	4.38
Presence of Children (%)	9.94	32.19	22.02	8.45	20.96
Number of Households	**825**	**11125**	**22252**	**11126**	**45328**

[1]: Age of female household heads was used for married-couple households.

CHAPTER 5

CONCLUSIONS AND DISCUSSIONS

A global epidemic of overweight and obesity coexists with undernutrition. Understanding how socio-economic, demographic and behavioral factors influence food choices is key to developing a solution to effectively promote healthy diet choices. This dissertation attempts to add efforts to this agenda from three different yet related perspectives.

The first essay proposes and evaluates a new modeling technique for multivariate binary responses data, a data feature of growing interest in the study of consumer behavior. This new estimator considers binary responses as being generated from a truncated multivariate discrete distribution. Specifically, the discrete normal probability mass function, which has support on all integers, is extended to a multivariate form. Truncating this point probability mass function below zero and above one results the multivariate binary discrete normal distribution. Compared to traditional MVP model, the discrete normal discrete normal estimator is easier for implementation in that MBDN estimators takes less time and runs into less problems during estimation. The MBDN estimator produces comparable marginal effects to the MVP. And the unconditional marginal effects under MBDN is straightforward and much easier to compute.

The second essay conducts an empirical investigation of household purchase decisions of tobacco and alcohol products in Poland. Harmful use of tobacco and alcohol is one of the main risk factors for adverse health and social consequences. Tobacco and alcohol use rates in Poland are high and their expenditures accounts for 7.6% of consumer expenditure. Such share is very high (ranking 8^{th}) compared to other countries. An analysis of the determinants of household

tobacco and alcohol consumption remains important in providing insights for reduce or prevent tobacco and alcohol use.

The second study in this dissertation takes advantage of household survey data collected by Poland's Main Statistical Office that is not publicly available. A pooled cross-sectional data of 77,043 observations in the period of 2005 to 2008 were used. A multivariate sample selection model is applied using full information maximum likelihood estimation. This censored system accommodates sample selection and correlation between the consumption of tobacco and alcohol.

Empirical investigation indicates that the decisions to smoke and drink as well as the associated expenditure levels are positively correlated. The empirical analysis revealed households features associated with tobacco and alcohol use and therefore helps to identify target for possible intervention.

The last essay of this dissertation studies at-home consumption of fresh produce in the Unites States. This study provides a holistic profile of fresh produce choices and expenditures, including expenditure on fresh produce, frequency of purchase, variety of selection, and use of deals and coupons. A profile of consumers by consumer group was developed using Nielsen Homescan panel in 2014. Socio-economic and demographic features indicating consumer disadvantage were given special attention. Low income, low education, very young household heads, and African American consumers are shown to be at higher risk of low and infrequent consumption of fresh vegetables and fruits. This study limits to at-home consumption of fresh vegetables and fruits. Processed produce such as frozen, canned and dried fruits and vegetables, garden fresh produce, and away-from-home foods were not included in the current analysis. It would be interesting for future research to include different components of fruit and vegetable

consumption and extend the comparison across various forms in which fruits and vegetables are purchased.

REFERENCES

Amemiya, T. 1985. Advanced Econometrics. Cambridge: Harvard University Press.

Bilgic, A. and S. T. Yen. 2015. "Household Alcohol and Tobacco Expenditures in turkey: A Sample-selection System Approach." *Contemporary Economic Policy* 33(3): 571-585.

Bettman, J.R. 1979. "Memory Factors in Consumer Choice: A Review." *Journal of Marketing* 43:37–53.

Blaylock, J.R., and W.N. Blisard. 1992. "U.S. Cigarette Consumption: The Case of Low-Income Women." *American Journal of Agricultural Economics* 74(1992): 698–705.

Bromley, R., and C. Thomas. 1993. "The Retail Revolution, the Careless Shopper and Disadvantage." *Transactions of the Institute of British Geographers* 18(2):222-236.

Cappellari, L. and S. P. Jenkins. 2003. "Multivariate Probit Regression Using Simulated Maximum Likelihood." The Stata Journal 3(3): 278-294.

Dong, D. and E. Leibtag. 2010. "Promoting Fruit and Vegetable Consumption: Are Coupons More Effective Than Pure Price Discounts?" *Economic Research Report* No. 96 (2010 June), ERS, USDA.

Economic Research Service (ERS), USDA. 2015. "Percent of consumer expenditure spend on food, alcoholic beverages, and tobacco that were consumed at home, by selected countries, 2014." Online available at http://www.ers.usda.gov/data-products/food-expenditures.aspx#26654 (Accessed Jul., 2016).

Economic Research Service (ERS), USDA. 2016. "Diet Quality & Nutrition". Online available at http://www.ers.usda.gov/topics/food-choices-health/diet-quality-nutrition/background.aspx (Accessed Jun., 2015).

Fair Trading Commission. 2016. "Looking Out for the Vulnerable Consumer." http://www.ftc.gov.bb/index.php?option=com_content&task=view&id=149&Itemid=28 (Accessed May, 2016.)

Fitzmaurice, G.M., N. M. Laird, and A.G. Rotnitzky. 1993. "Regression Models for Discrete Longitudinal Responses." *Statistical Science* 8(3): 284-299.

Florkowski, W. J. and K. T. McNamara. 1992. "Policy Implications of Alcohol and Tobacco Consumption in Poland." *Journal of Policy Modeling* 14(1): 93-98.

Gensch, D.H. 1987. "A Two Stage Disaggregate Attribute Choice Model." *Marketing Science* 6 (Summer 1987):223–231.

Geweke, J. 1989. "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica* 57(6):1317-1339.

Giskes, K., G. Turrell, C. Patterson, B. Newman. 2002. "Socioeconomic Differences among Australian Adluts in Consumption of Fruit and Vegetables and Intakes of Vitamin A, C, and Folate." *Journal of Human Nutrition and Dietetics* 15: 375-385.

Greene, W., 2009. "Discrete Choice Modeling" in: Mills, T.C., Patterson, K. (Eds.), Applied Econometrics. Handbook of Econometrics, Vol.2. Palgrave Macmillan, London.

Guthrie, J., B. Lin, J. Reed, and H. Stewart. 2005. "Understanding Economic and Behavioral Influences on Fruit and Vegetable Choices." *Amber Waves* 3.2 (Apr., 2005):36-41.

Herman, D., G. Harrison, A. Afifi, and E. Jenks. 2008. "Effect of a Targeted Subsidy on Intake of Fruits and Vegetables among Low-Income Women in the Special Supplemental Nutrition Program for Women, Infants, and Children." *American Journal of Public Health*: (January 2008) 98(1): 98-105.

Freeth, S. *Smoking-Related Behaviour and Attitudes 1997: A Report on Research Using the Omnibus Survey Produced on Behalf of the Department of Health.* London: Office for National Statistics, 1998.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52(3):681-700.

Hajivassiliou, V. A. and D. L. McFadden. 1998. "The Method of Simulated Scores for the Estimation of LDV Models." *Econometrica* 66(4): 863-896.

Heien, D. and C.R. Wessells. 1990. "Demand systems Estimation with Microdata: A censored Regression Approach." *Journal of Business & Economic Statistics* 8(3): 365-371.

Heckman J.J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Department Variables and a Simple Estimator for Such Models." *Annals of Economic Social Measurement* 5 (4):475–492.

_____. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1) (Jan., 1979):153–161.

Jones, A.M., and J.M. Labeaga. 2003. "Individual Heterogeneity and Censoring in Panel Data Estimates of Tobacco Expenditure." *Journal of Applied Econometrics* 18(2003):157–77.

Keane, M. P. 1994. "A Computationally Practical Simulation Estimator for Panel Data." *Econometrica* 62 (1):95-116.

Kemp, A. W. 1997. “Characterizations of a Discrete Normal Distribution.” *Journal of Statistical Planning and Inference* 63 (2):223-229.

Lee, L-F., and M.M. Pitt. 1986. “Microeconometric Demand Systems with Binding Nonnegativity Constraints: The Dual Approach.” *Econometrica* 54:1237–42.

Lindsey, J.K. 1996. *Parametric Statistical Inference*. New York: Oxford Science Publications.

Lin, B.H., J. Guthrie, and J.R. Blaylock. 1996. “The Diets of America’s Children: Influences of Dining Out, Household Characteristics, and Nutrition Knowledge.” U.S. Dept. Agr., Econ. Res. Serv., AER-746, Dec. 1996.

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

MATLAB and Statistics Toolbox Release 2014b, The MathWorks, Inc., Massachusetts, United States.

McFadden, D. 1980. “Econometric Models for Probabilistic Choice among Products.” *Journal of Business* 53(3) Part 2: Interfaces between Marketing and Economics 13-29.

Mullahy, J. 2011. “Marginal Effects in Multivariate Probit and Kindred Discrete and Count Outcome Models, with Applications in Health Economics.” National Bureau of Economic Research (NBER) working paper 17588.

Pearson K. 1900. “Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable.” *Philosophical Transactions of the Royal Society of* London, Series A, Containing Papers of a Mathematical or Physical Character: 195:1-47.

Perali, F., and J.P. Chavas. 2000. “Estimation of Censored Demand Equations from Large Cross-Section Data.” *American Journal of Agricultural Economics* 82:1022–37.

Pierani, P. and S. Tiezzi. 2009. “Addiction and Interaction between Alcohol and Tobacco Consumption.” *Empirical Economics* 37(1): 1-23.

Prentice, R.L. and L.P., Zhao. 1991. “Estimating Equation for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses.” *Biometrics* 47 (3): 825-839.

Pudney, S. 1989. “Modelling Individual Choice: The Econometrics of Corners, Kinks, and Holes.” Cambridge, UK: Blackwell Publishers.

Rosinski, J., and S.T. Yen. 2004. “A Note on the Conditional Moments of Limited Dependent Variable Models with a Transformed Dependent Variable.” Unpublished, Dept. Agr. Econ., The University of Tennessee, Knoxville, 2004.

Shocker A.D., M. Ben-Akiva, B. Boccara, and P. Nedungadi. 1991. "Consideration Set Influences on Consumer decision-Making and Choice: Issues, Models and Suggestions." *Marketing Letters*, 2(Aug., 1991):18–197.

Shonkwiler, J.S., and S.T. Yen. 1999. "Two-Step Estimation of a Censored System of Equations." *American Journal of Agricultural Economics* 81:972–82.

Spanos, A. 1999. Probability Theory and Statistical Inference: Econometric Modeling with Observational Data. Cambridge, UK: Cambridge University Press, 1999.

Stewart, H. and S.T. Yen. 2004. "Changing household characteristics and the away-from-home food market: a censored equation system approach." *Food Policy* 29(6): 643–658.

Wales, T.J., and A.D. Woodland. 1983. "Estimation of Consumer Demand Systems with Binding Non-negativity Constraints." *Journal of Econometrics* 21:263–85.

Wright P. and F. Barbour. 1977. "Phased Decision Strategies: Sequels to Initial Screening." in *Multiple Criteria Decision Making: North Holland TIMS Studies in management Science*, M. Starr and M. Zeleny, eds. Amsterdam: North-Holland Publishing Company, 91 – 109.

World Health Organization (WHO). 2016a. "Controlling the global obesity epidemic." Online available at http://www.who.int/nutrition/topics/obesity/en/ (Accessed Jul., 2016).

____. 2016b. "Healthy diet" Online available at http://www.who.int/mediacentre/factsheets/fs394/en/ (Accessed Jul., 2015).

____. 2016c. *Health topics: Tobacco*. Online available at http://www.who.int/topics/tobacco/en/ (Accessed Jan., 2016).

____. 2016d. *Health topics: Alcohol*. Online available at http://www.who.int/topics/alcohol_drinking/en/ (Accessed Jan., 2016).

____. 2016e. *WHO Report on the Global Tobacco Epidemic, 2015 – Country profile: Poland.* http://www.who.int/tobacco/surveillance/policy/country_profile/pol.pdf?ua=1(Accessed Jan., 2016).

____. 2016f. *WHO Global status report on alcohol and health 2014 – country profile: Poland.* Online available at http://www.who.int/substance_abuse/publications/global_alcohol_report/profiles/pol.pdf?ua=1 (Accessed Jan., 2016).

Yen, S. 2005. "A Multivariate Sample-selection Model: Estimating Cigarette and Alcohol Demands with Zero Observations." *American Journal of Agricultural Economics* 87(2) (May, 2005): 453-466.

APPENDIX

Proposition: Matrix **A** in Equation (2.18) is a zero matrix when the regressors are identical across equations.

Proof:

When regressors are identical across equations, the concatenated matrix of explanatory variables can be written as a kronecker product as follows: $\boldsymbol{X_i^* = I_M \otimes x_i^*}$. Equation (1.19) then becomes

(A1) $\frac{1}{n}\sum_i(\boldsymbol{y_i \Sigma^{-1} I_M \otimes x_i^*}) = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_i E(\boldsymbol{y_i})\boldsymbol{\Sigma^{-1} I_M \otimes x_i^*}$,

which can be further re-organized as follows:

(A2) $\boldsymbol{\Sigma^{-1}\tilde{y} = 0}$,

where $\tilde{\boldsymbol{y}} = \left[\frac{1}{n}\sum_i(y_{1i} - E(y_{1i}))\boldsymbol{x_i^*}\,;\quad \frac{1}{n}\sum_i(y_{2i} - E(y_{2i}))\boldsymbol{x_i^*}\,;\quad \ldots\quad \frac{1}{n}\sum_i(y_{Mi} - E(y_{Mi}))\boldsymbol{x_i^*}\right]$ is a M x k matrix, where k is the number of explanatory variables. The elements of $\tilde{\boldsymbol{y}}$ can be solved by Cramer's Rule. The system in Equation (A2) can be broke down into k systems, where the to-be-solved vector is a column of $\tilde{\boldsymbol{y}}$ and the answer vector is the corresponding zero column vector of the M x k answer matrix **0**. For each system, let D=det($\boldsymbol{\Sigma^{-1}}$), and D_m be the coefficient determinant with answer-column values in the m^{th} column of the coefficient matrix $\boldsymbol{\Sigma^{-1}}$. Since the answer-column is a zero vector, $D_m = 0\ \forall$ m=1, 2, … M. This solution holds for all k systems and thus the elements of $\tilde{\boldsymbol{y}}$ are zeros. Above result expressed in general notation is as follows:

(A3) $\frac{1}{n}\sum_{i=1}^{n}\left(y_{mi} - E(y_{mi})\right)\boldsymbol{x_i^*} = 0;\ \forall \mathrm{m} = 1, 2, \ldots \mathrm{M}.$

An examination of the extra component, A, in Equation (2.18) is as follows:

$$\boldsymbol{A} = \frac{1}{n}\sum_{\boldsymbol{i}} \boldsymbol{X_i \beta}\left(\boldsymbol{y_i} - \boldsymbol{E(y_i)}\right)$$

$$= \frac{1}{n}\sum_{\boldsymbol{i}} \begin{bmatrix} 1 & \boldsymbol{x_{1i}} & 0 & \boldsymbol{0} & \cdots & 0 & \boldsymbol{0} \\ 0 & \boldsymbol{0} & 1 & \boldsymbol{x_{2i}} & \cdots & 0 & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \boldsymbol{0} & 0 & \boldsymbol{0} & \cdots & 1 & \boldsymbol{x_{2M}} \end{bmatrix} \begin{bmatrix} \beta_{10} \\ \boldsymbol{\beta_{11}} \\ \beta_{20} \\ \boldsymbol{\beta_{21}} \\ \vdots \\ \beta_{M0} \\ \boldsymbol{\beta_{M1}} \end{bmatrix} [y_{1i} - E(y_{1i}) \quad y_{2i} - E(y_{2i}) \quad \cdots \quad y_{Mi} - E(y_{Mi})]$$

where β_{m0} is the parameter associated with the constant term for the m^{th} response variable, and $\boldsymbol{\beta_{m1}}$ is a vector of parameters of the explanatory variables for equation m (m=1, 2, … M). The (m, k)th element of matrix **A**, a_{mk}, is as follows:

(A4) $$a_{mk} = \beta_{m0}\frac{1}{n}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(y_{ki} - E(y_{ki})\right) + \frac{1}{n}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(y_{ki} - E(y_{ki})\right)\boldsymbol{x_{mi}^{*}\,\beta_{m1}},\ m, k = 1,2, \ldots M.$$

The first term of the RHS is zero under the first moment condition, regardless whether the regressors are identical across equations. The second term reduces to zero when there are identical regressors (Equation (A3) as previously proven). Therefore the extra component in Equation (2.18) is a zero matrix and thus the second moment equation is the maximum likelihood equation under identical regressors. The combination of the first two moment equations leads to the result that the estimated correlation among response variables under the MBDN model matches the sample correlation.