2018

Excess Zeros, Endogenous Binary Indicators, And Self-Selection Bias With Application To First Marriage, Smoking And Drinking Outcomes

Lateef Subair
University of Mississippi

EXCESS ZEROS, ENDOGENOUS BINARY INDICATORS, AND SELF-SELECTION BIAS WITH APPLICATION TO FIRST MARRIAGE, SMOKING AND DRINKING OUTCOMES

A Thesis
presented in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Economics
The University of Mississippi

by

LATEEF A. SUBAIR

May 2018

Committee in Charge:

Professor Walt Mayer, Chair
Professor Mark Van Boening
Associate Professor William Chappell
Professor Yi Yang

ABSTRACT

This dissertation examines empirical application of the zero-inflated ordered probit (ZIOP) model to the impact of first marriages on smoking and alcoholic beverage consumption. The data for this study is drawn from the National Longitudinal Survey of Youth (1997). In my ZIOP model analysis of the impact of first marriage on smoking and alcoholic consumption, I juxtaposed the ZIOP model with popular models in health economics literature like the ordered probit (OP) model, the ordered probit endogenous dummy (OP-ED) model, the zero-inflated ordered probit model correlated (ZIOPC) and the Heckman sample selection ordered probit (SSOP) model. The analysis highlighted four sets of result. First, all the statistical tests of the model specifications, including the Vuong test, and information criteria, show that the ZIOP model of the impact of marriage on smoking and alcoholic beverage consumption is superior to the OP, OP-ED, SSOP, and ZIOPC models. Second, first marriages increase the probability of zero consumption of tobacco products and alcoholic beverages. Third, conditional on participation, the probability of zero alcohol consumption is not significantly different from zero. The converse is true for the smoking sample. Last, the benefits of first marriage in terms of reduced smoking and drinking is diminishing in the ordinal levels of the intensities of tobacco and alcoholic beverage consumption.

DEDICATION

This dissertation is dedicated to my late father, Subair Alade Oladipupo and my mother Alhaja Munirat Titilayo Subair, for striving to get me educated, and to my fiancée, Idowu-Subair Rukayat Ronke for standing by me and encouraging me in all my endeavors.

LIST OF ABBREVIATIONS AND SYMBOLS

AIC	Akaike information criteria
BIC	Bayesian information criteria
CDC	Centers for Disease Control
DGPs	Data generating processes
FRED	Federal Reserve Economic Data
GHQ	Gauss-Hermite quadrature
GHK	Geweke-Hajivassiliou-Keane
MSL	Maximum simulated likelihood
NDCP	Office of National Drug Control Policy
NLSY97	National longitudinal survey of youth 1997
OP	Ordered probit
OP-ED	Ordered probit endogenous dummy
ZIOP	Zero-inflated ordered probit
ZIOPC	Zero-inflated ordered probit correlated
SSOP	Heckman Sample Selection Ordered Probit Model

ACKNOWLEDGEMENTS

The analyses in this dissertation benefit greatly from the advice of my thesis committee members: Professors Walt Mayer (Chair), Mark Van Boening, Bill William Chappell and Yi Yang.

I am greatly indebted to Professor Bill William Chappell for his fatherly advice and mentorship spanning many years at this university.

I am indebted to Professor Walt Mayer, the chair of my dissertation committee, for his indispensable advice on this dissertation. I am greatly indebted to Professor Mayer for taking his time to help me with suggestions that greatly improve the quality of this dissertation. I am also indebted to Professor Mayer for his academic advice and prompt recommendation letters.

I am also grateful to Professor Mark Van Boening for his invaluable advice on this dissertation. I cannot thank Professor Van Boening enough for helping me critique this dissertation and for offering invaluable advice on the content of this dissertation.

I also thank Professor Yi Yang for agreeing to serve on my committee.

Finally, I am grateful to my family members, loved ones, and my friends for their love and support.

Any error and conclusions in this dissertation are my own.

TABLE OF CONTENTS

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Researchers in medical and social sciences have long established that tobacco users are more likely to die from heart disease, lung cancer and other fatal ailments than non-users of tobacco products. The World Health Organization (2015) estimates that smoking accounts for 6 million deaths annually around the world, including some 600,000 people who die from passive or second-hand smoking. In the United States alone, more than 16 million people suffer from diseases caused by smoking (U.S. Department of Health and Human Service, 2014). Several studies have also shown that excessive consumption of alcohol contributes to health-related problems (See Duncan et al., 2006; and Leonard et al., 2014). Also, in recent years, trends in smoking among young adults' population, including young females, have been on the rise (Harris and Zhao, 2007). Because of these recent trends in global and smoking habits, policy makers and governments all over the world are devoting more resources aimed at stemming smoking and excessive consumption of alcohol.

Despite the enormous resources devoted to stemming smoking and excessive alcohol consumption, policy makers do not often consider the role that marriage plays in reducing smoking and drinking. Since the seminar series and publications of Gary Becker (1972, 1973) on marriage markets, the marriage institution has been studied as a market whose participants are economic agents seeking utility maximization. Later studies inspired by Becker's (1973, 1974) surplus benefits theory of marriages show that individuals who had experienced first marriages are less likely to be involved in health-related risky behaviors, including smoking and drinking (See Leonard et al., 2014; Duncan et al., 2006).

In my dissertation, I focus on the impact of first marriage on the probabilities of zero consumption of alcoholic beverages and tobacco products. However, I also consider the probabilities of consumption at other positive levels. I focus on the endogenous participation problem with regards to estimating the effect of first marriage on zero consumption of alcoholic beverages and tobacco products. Endogenous participation entails a regime split between participation and consumption decisions. The split between participation and consumption decisions in health economics literature is common in modeling discrete choice responses of durable and addictive goods. To specify the endogenous participation problem, I apply the zero-inflated ordered probit (ZIOP) model, a model developed by Harris and Zhao (2007). For comparative analysis and goodness of fit measures, I also juxtapose the ZIOP model with similar models.

Endogenous participation estimation issues arise from the differences between the decision to participate in an activity (the participation decision) and the intensity of participation conditional on the decision to participate in that activity (the consumption decision). Take consumption of tobacco products as an example. Users of tobacco product must first decide whether to smoke or not. Then conditional on the decision to smoke, these users will then decide how much tobacco products to consume. *How much to consume* in this case includes zero consumption. Thus, there are two sources of zero consumption in this example: outright abstention from smoking and occasional/infrequent smoking. Because of timing and resources constraints, outright abstention (or nonparticipation) and infrequent/occasional consumption (or zero consumption conditional on participation) are often lumped together as zero consumption values in surveys, but these zeros have separate data generating processes (DGP). At the heart of endogenous participation problem in discrete choice response modeling is the treatment of these zeros in surveys of addictive and discrete goods.

With regards to discrete choice responses, modeling these separate sources of zeros in surveys of addictive goods is a matter of differentiating between three types of zero. These zeros are 'abstention' zeros, corner solution zeros, and excess zeros. What are the differences between these three sources of zero? What are the specification procedures of modeling them? What differences would the application of any of these zero consumption models make with respect to estimating the impact of first marriage on the probabilities of drinking and smoking? And is there any efficient model other than the ZIOP model in this regard? These are some of the questions I want to answer in this dissertation. A detailed explanation of each type of zero follows from here.

Models of demand for alcohol and tobacco products have attracted the attention of economists since early 1960s.[1] By design, most studies on addictive products like tobacco utilize sampling frameworks in which the zero outcome or consumption of these products are coded with the number zero (0). In this sense, using the number zero (0) and consecutive ordinal numbers for positive consumption levels is conventional in economic surveys. I follow this convention starting by coding the consumption levels as 0, 1, 2, or 3.

There are different estimation strategies of modeling each of the three types zero outcomes. However, the choice of an estimation method depends on *why* the zeros are in the survey. First, with regards to the first type of zeros, some respondents optimally chose to consume zero amount of a good. In this case, these zero outcomes index demand for a perfectly inelastic good. I call these zeros 'abstention' zeros. Single equation models like the ordered probit (OP) model and the ordered logit model are often used as estimation strategies when the dataset include ordinal measures of zero and positive levels of consumptions as the dependent variable. I focus on the OP model for this type of zeros in this study. The predictor variables for the OP model of addictive

[1] Cragg (1971) discusses the early works on limited dependent variables.

goods like tobacco products are usually demographic variables like age, gender, race, marital status, and so on.

With regards to modeling addictive good, the OP model has one major drawback. The OP assumes that all the observation units are participants. That is, the OP model does not condition consumption on participation. For example, in this instance the OP model treats the probability of observing a positive value $P(y > 0)$, and the probability of observing an actual value given that it is positive, i.e $P(y|y > 0)$, as being determined by the same data generating process.

In this study, I include four demand-related variables in the OP model in addition to the socio-economic predictors of tobacco and alcoholic beverage consumption. These economic variables are own-price, income, the price of substitutes and the price of complements. I use these variables to test the impact of own-price, income, substitutes and complements on tobacco and alcoholic beverage consumptions. These four predictors provide additional predictive power to the OP model. For example, some smokers who have not smoked recently because the price of the product is relatively high may consider consuming some positive amounts of cigarette if there is a fall in the price. This hypothesis cannot be tested if the own-price variable is not included in the OP model.

Cases of "abstention" zero responses in surveys abound in health economics literature. Several studies and surveys have shown that, perhaps due to health implications of risky behaviors, a clear majority of adult population abstain from smoking. For this proportion of the adult population, prices of cigarette and other factors affecting the demand for tobacco products will not change their demand from zeros to some positive demands.

One can also consider an extension of the OP model to account for endogenous predictors. The ordered probit model with endogenous dummy (OP-ED) model, a combination of an OP

model and a binary discrete model, can be used for this purpose. I consider this extension in this study. Specifically, I use the OP-ED model to model the possibility of bias and selectivity into marriage. I adopt the maximum simulated likelihood (MSL) estimation strategy to estimate the OP-ED model (See Train, 2003; Roodman, 2013; and Green and Hensher, 2010).

The second type of zero in surveys of addictive goods is associated with "corner solution" models. For corner solution models, zero consumption is an optimal solution to some economic agents' utility maximization problems. But some economic agents may choose to consume positive amounts of the addictive good if certain economic variables change. The corner solution model is a Tobit model which features both discrete and continuous dependent variables. For positive response values such as expenditure on positive amounts of tobacco products, corner solution model features continuous variables, otherwise the response variable 'piles up' at zero, discrete variable. The corner solution model is a censored regression model (Wooldridge, 2010). Like the OP model, the major drawback of the Tobit model is that the probability of observing a positive value, $P(y > 0)$, and the probability of observing an actual value given that it is positive, $P(y|y > 0)$ are determined by the same data generating process. That is, just like the OP model, corner solution models like the Tobit model treats all observation units as participants. In any case, the application of the corner solution model is outside the scope of this study. Thus, I will not analyze corner solution models in this dissertation.

The third type of zeros are known as "excess" zeros in the economics literature. Excess zero models are often associated with survey questions on addictive and durable goods over a short period. This short timeframe can lead to regime split between participation decisions and consumption decisions. For example, respondents in the National Longitudinal Survey of Youth 1997 (NLSY97) were asked some questions about their smoking habits. Two of these questions

are: "*During the last 30 days, how many days did you smoke a cigarette*? and "*When you smoke during the last 30 days, how many cigarettes did you usually smoke each day*? Since these NLSY97 questions provide snapshots of the demand for tobacco products within a very short period (30 days), it might be difficult to distinguish between respondents who have *never* smoked (coded with a "0" on the NLSY97) and infrequent smokers who have not smoked within the short period of the survey question (also coded with a "0" in the NLSY97). Thus, at the zero consumption levels, the answers to the NLSY97 questions above comprise those respondents whose demand for cigarettes is perfectly inelastic ("abstention" zeros) smokers who may consider switching to positive consumption if the conditions are right ("corner" solution zeros), and infrequent smokers.

In the third case of zero consumption above, a respondent must first decide whether to be a smoker. Conditional on participation, he must decide the number of positive cigarettes to consume. But the survey timeframe may be too short (usually 30 days) to capture this behavior. A respondent who smokes occasionally may chose zero amount of cigarette in the NLSY97 because he has not smoked in the last 30 days. "Genuine" zero and corner solution models like the OP model cannot adequately describe this optimal consumption choice. A model that captures the separate DGP that characterizes occasional or infrequent consumption habits are the so-called zero inflated model. As an inflated zero model, the ZIOP model is a split between participation and consumption behavior. Thus, at the level of zero consumption, the ZIOP model is a double hurdle model of outright abstention (nonparticipation decision) and participation with zero consumption.

If the underlying DGP is different for both consumption and participation decisions, models of zero consumptions that do not specify an inflated zero model may lead to inconsistent estimates. Following Harris and Zhao (2007), I demonstrate that traditional ordered probit models cannot

distinguish between zeros due to outright abstentions and those due to participation with zero consumption. Since the ZIOP model is a regime split that combines the probit and ordered probit models, it specifies both the participation and consumption decisions to model inflated zeros. Apart from the ZIOP model, another prominent zero-inflated double hurdle model in econometrics literature is the zero-inflation Poisson model (Lambert, 1992). Double hurdle zero-inflated models are built on earlier works of Tobin (1958) and Cragg (1971)[2] hurdle models. Figure 1.1 below presents a summary of the three types of zero and their corresponding models.

The National Longitudinal Survey of Youth 1997 (NLSY97) is the main source of data for this study. The NLSY97 is a nationally representative sample of 8,984 respondents. Apart from questions about smoking and drinking habits of the respondents, the NLSY97 also includes questions about the years of first marriages and other demographic information of the respondents. I use these variables to model the impact of first marriage on smoking and drinking habits of the respondents. To capture the impact of marriage on smoking and drinking before and after the year of first marriage, I specify a spline regression for all the models in this study.

Apart from the traditional OP and OP-ED models, I also consider the ordered probit model with the sample selection (SSOP) model. Like De Luca and Perotti (2011), I use the SSOP model to address the problem of sample attrition. As of 2014, the NLSY (97) has recorded an attrition rate of about 20%. Thus, to determine the extent for survivorship bias in NLSY97 dataset, I include a specification test for the SSOP model in Chapter 6.

[2] Tobin (1958) proposes the Tobit model while Cragg (1971) proposes the double hurdle model. These two models are the early works on censored limited dependent variable models.

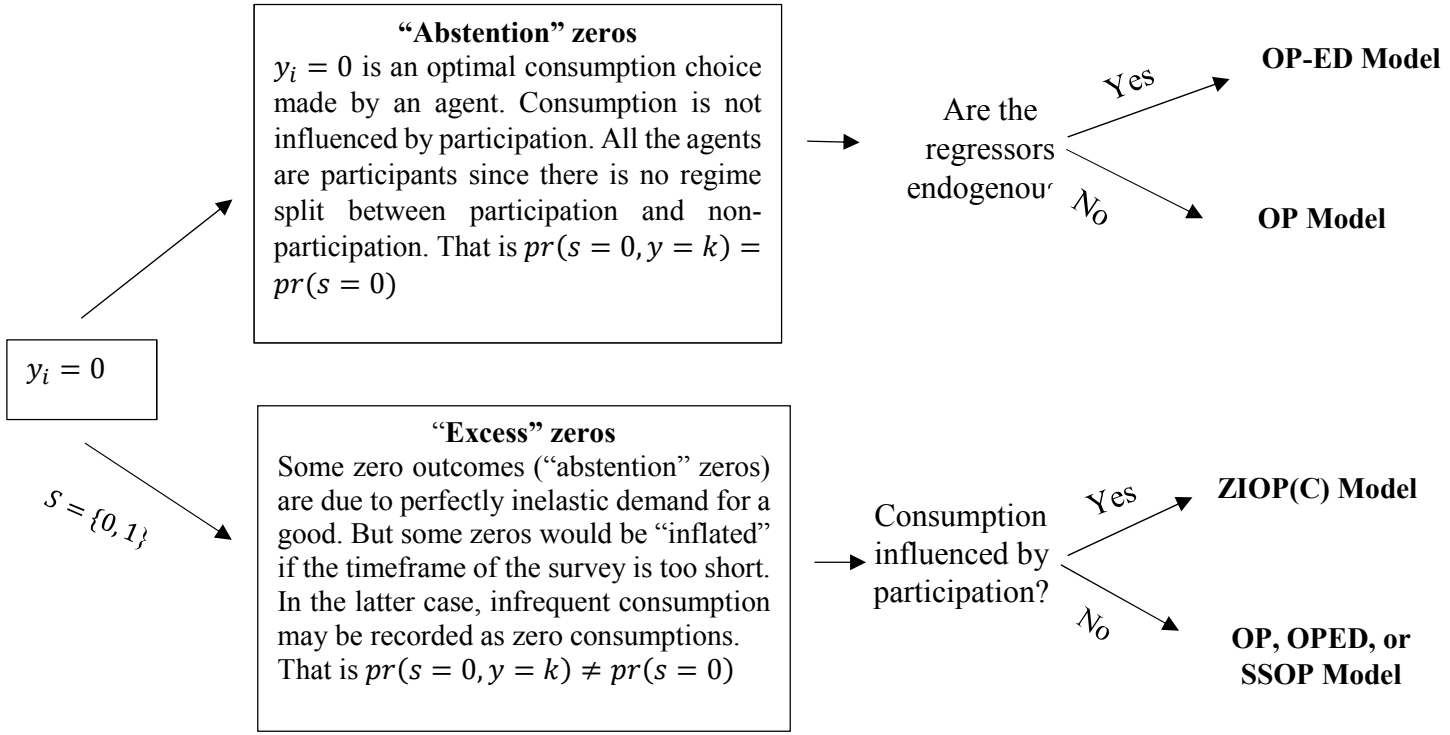

Figure 1.1: Types of Zero, and their Corresponding Models

Finally, to determine the ultimate model that best fits the data, I juxtapose the ZIOP model with the OP, OP-ED and SSOP models. I run goodness of fit tests between model pairs based on the research questions that I raised earlier. These model pairs and their corresponding research questions (with emphasis on zero outcomes) are shown in Figure 1.2 below. The selection criteria in these horseraces are examined in Chapter 6.

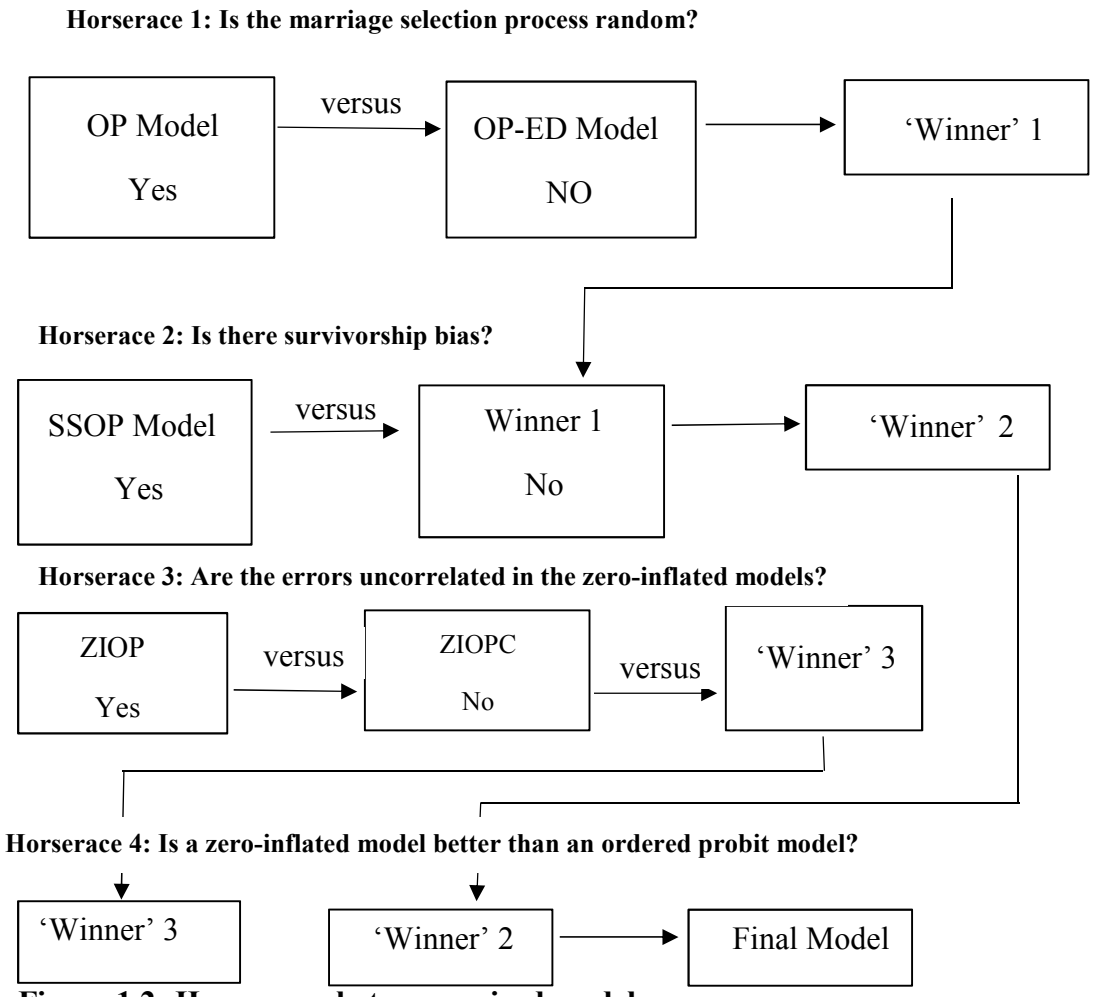

Figure 1.2: Horseraces between paired models

Using the rival estimation methods to the ZIOP model to specify inflated zeros may lead to wrong or sub-optimal policy formulation. To some degree, the methodology and procedure of testing the goodness of fit between rival models depend on whether the models are nested or non-nested. Tests of rival nested models use straightforward and familiar testing procedures, including

the t-test, the chi-square test, and information criteria like the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), and the Vuong (1989) test. The Vuong (1989) test is designed for testing the goodness of fit of non-nested models.

Beyond specifying double hurdle models and single equation models to capture the effects of first marriage on alcohol and tobacco consumptions, I also address the issue of self-selection in the first marriage search process. The transition from a single to a married life may be driven by self-selection. It is conceivable that married individuals may deliberately choose to live healthier lives to improve their chances of finding a spouse. If this is the case, the marriage market is not entirely driven by a random process, and this is a major violation of the conditional expectation of first marriage in the consumption decision equation's error term. I test the hypothesis of first marriage self-selection in this study.

I address the self-selection problem in first marriages by estimating an ordered probit with an endogenous marriage dummy variable. I refer to this model as the ordered probit endogenous dummy (OP-ED) model. The OP-ED model is estimated by maximum simulated likelihood (MSL) method. The first stage of the OP-ED model estimates a probit model of first marriage, while the second stage estimates an ordered-probit model for the effect of first marriage on the probabilities of smoking and drinking. The specification, method, identification and other estimation issues related to the OP-ED model can be found in the Train (2003), Bratti and Miranda (2010), Roodman (2013), and Green and Hensher (2010).

To the best of my knowledge, this study is the first to model the benefit of first marriage on abstentions from smoking and drinking as a double hurdle model of excess zeros. Similar studies on the impact of marriage on smoking and drinking behavior include Duncan (2006), Ali and Ajilore (2010), Murray (2000) and, Leonard et al. (2014). In summary, I model the influence

of first marriage on drinking and smoking, with emphasis on the probability of occurrence of inflated or excess zero consumption in this study. Acceptable models of excess zeros models of smoking and drinking in the economics literature include the ZIOP model, a double hurdle model which treats the underlying DGP of the zero outcomes in two interrelated steps.

First, individuals must first decide whether to participate. This step is the participation decision. Second, conditional on participation, individuals must decide *how much* of the addictive good to consume, including zero consumption. This second step is known as the consumption decision. At least one of the SSOP, the OP, and the OP-ED models should fit the data in this case, and I use selection criteria to determine this ultimate model. I also use the maximum simulated likelihood (MSL) model to specify the endogeneity issues in the marriage search process.

The key findings of my dissertation are:

1) Goodness of fit tests show that the ZIOP model is a better fit for the underlying dataset than the OP, the OP-ED, and the SSOP models.

2) First marriage increases the probability of zero alcoholic beverage consumption in the three years preceding the year of first marriage (hereinafter called "around the year of first marriage" period) and in the five years following the first marriage event (hereinafter called "after the year of first marriage" period).

3) First marriage increases the probability of zero tobacco consumption "around the year of first marriage" and "after the year of first marriage".

4) In the ZIOP model, the estimated probability of zero alcoholic beverage consumption conditional on participation is zero. In contrast, for most predictors, the estimated probability of zero consumption of tobacco product conditional on participation

is large and statistically different from zero. These results imply that while some smokers are infrequent/occasional smokers, most drinkers are social drinkers.

5) The benefits of first marriage in terms reduced alcoholic and tobacco consumption diminished in smoking and drinking intensity levels. The marginal benefits of the impact of marriage on drinking and smoking peaked at the ordinal levels of one and two for the drinking and smoking sample respectively.

6) In the three years preceding the first marriage event and the five years following the first marriage event, males have higher probabilities of zero consumption of alcoholic beverage and tobacco products than females. That is, males benefit more than females in terms of reduced smoking and drinking.

The remainder of this dissertation is organized as follows. Chapter 2 presents the literature review, and chapter 3 describes the data and descriptive statistics. Chapter 4 discusses the methodology. Chapter 5 presents the econometric specifications. Chapter 6 presents the model selection criteria and results. Chapter 7 discusses the estimated results. Chapter 8 addresses the policy implications. Chapter 9 concludes this dissertation.

CHAPTER 2: LITERATURE REVIEW

In this Chapter, I review studies related to my dissertation. These include the literatures on the influence of marriage on drinking and smoking, past studies on modelling of excess zeros in econometrics, and the treatment of self-selection in marriage models. I begin this chapter by reviewing the literature on the theoretical benefits of marriage. In this regard, the seminar series and publications of Garry Becker (1972, 1973) and the works of Matouschek and Rasul (2008) will be examined. Specifically, the theories of Becker (1973,1974), and Matouschek and Rasul (2008) are adapted to my dissertation to show that individuals get married because they get benefits that are associated with marriage. These benefits are realizable largely due to cooperation between the parties involved in a marriage. Reduced smoking and drinking are two of these benefits. After reviewing the theoretical literature of the benefits of marriage, I narrow my review to empirical studies on the benefits of marriage, including specific studies on the effect of marriage on smoking and drinking. Finally, I review the literature of econometric specification of smoking and drinking in this chapter.

2.1 Cooperation, monitoring, divorce propensities and benefits of marriage

The seminar series and publications of Becker (1973, 1974) are some of the earliest works on the marriage market. Beker regards the institution of marriage as a market because individuals compete when seeking mates. Becker (1973, 1974) posits that individuals derive surplus benefits from getting married. Further, Becker points out that economic theory is applicable to marriage because couples expect their post-marriage utilities from the surplus benefit of marriage to exceed

their pre-marriage utilities. The surplus benefits include specialization, economies of scale, and risk sharing in managing the home (Becker, 1974). Later studies on the nexus between marriage and health-related benefits frequently make use of Becker's theory on the marriage market. Like past studies, one of my major premises is that married individuals get health-related benefits.

In their paper titled "Economics of the Marriage Contract: Theories and Evidence" Matouschek and Rasul (2008) extend Becker work by focusing on marriage contracts. They show that the size of the surplus benefits or payoffs from marriage depend on constant monitoring and commitment between married couples. Matouschek and Rasul (2008) also model some costs that are incurred at the time of divorce or separation. However, since I am primarily concerned with analyzing the health-related benefits of first marriage with respect to zero probabilities of smoking and drinking, I am going to focus on the commitment and constant monitoring aspects of marriage economics. Hence, I present the model of the benefits from marriage.

2.1.1 Discounted benefits of marriage

Assume that at time $t = 0$, there is a unit mass pool of men and women who are old enough for marriage, and these individuals can live *ad infintum*. Individuals in the mass pool are faced with two choices: marry or remain single. At $t = 0$, all the individuals in the pool have perfect knowledge about the costs and benefits of marriage. Let b denote the benefits of marriage. These benefits include living a healthy life style through constant spousal monitoring, among others. However, these benefits are not realized until $t = 1$. Singlehood has no cost, but it has some benefits. Let s denote the payoff for individuals who remain in the single pool. The payoff for individual who returns to the single pool after divorce is also s, where $s \in [0, \infty)$. Matouschek and Rasul (2008) call $s \in [0, \infty)$ the outside option. Following Matouschek and Rasul (2008), I

assume that b is drawn from a distribution with cumulative distribution function $H(b)$, with limits $[0, \infty)$. Payoff s is randomly drawn from the cumulative distribution $F(s)$.

Beginning at $t = 1$, each partner in a marriage realizes $B + b$, where $B > 0$ and it is a fixed exogenous benefit of marriage. For example, B can be prestige and respect that married couples get from the society simply for being married. These payoffs are the same for men and women. After $t = 1$, married couples can choose to divorce or stay married. If the agents choose divorce or break up, the game ends. For married individuals, the cost of divorce (e.g., legal fees) is γ per partner. After divorce, each person in a marriage relationship realize s, the outside option. All periods $t = 2,3, \ldots$ are identical to $t = 1$. Following Matouschek and Rasul (2008), I am going to assume that each agent has a time discount factor, r and $r \in [0,1)$.

For any period $t > 0$, the expected payoff for each partner in a marriage is:

$$V_m = B + b + \int_0^{rV_m+\gamma} rV_m dF(s) + \int_{rV_m+\gamma}^{\infty} (s - \gamma) dF(s) \qquad (2.1)$$

In the second term on the right-hand side of equation (1), b is the benefit each partner gets for being married. The third term on the right-hand side occurs if $s < rV_m + \gamma$, implying that the payoff for remaining or returning to the single pool is less than the discounted payoff from marriage. Married couples in this situation would stay married. I also assume that $B > 0$ and $b > 0$ in this case. The last term on the right-hand side in equation 1 above occurs when $s \geq rV_m + \gamma$, which is when married couples would divorce because the cost of remaining or returning to the pool is less than the benefit of staying married. $B = 0$ and $b = 0$ in this case. Thus, a large value of b makes it less likely for the partners in marriage contracts to break up. A small b makes it more likely that a break up will occur.

2.1.2 Divorce propensities

Following Matouschek and Rasul (2008), individuals in marriage relationships review their payoffs at the end of each period and decide whether to stay in the relationship or leave. For married couples, the payoff after divorce is $(s-\gamma)$ while the payoff for remaining married is rV_m. Since married couples get divorced when $(s-\gamma) \geq rV_m$, the probability of getting a divorce in period 1 is $F(rV_m+\gamma)$, while the probability of staying married at the end of period 1 is $1-F(rV_m+\gamma)$. The probability of divorce at the end of period 2 is:

$$\big(1-F(rV_m+\gamma)\big)\times\big(F(rV_m+\gamma)\big) \tag{2.2}$$

Thus, the probability of divorce at the end of period $t=n$ is:

$$\big(1-F(rV_m+\gamma)\big)^{n-1}\times\big(F(rV_m+\gamma)\big) \tag{2.3}$$

2.1.3 Benefits from marriage: The monitoring and cooperative game

In this section, I develop the theoretical framework of monitoring and cooperation among married couples. Maintaining healthy life styles in marriages involve constant monitoring by the parties involved, as each partner can either cooperate or not. At the start of the period $t>0$, let b denote the benefits from marriage. Following Matouschek and Rasul (2008), I simplify the model by assuming that in this the exogenous benefit of marriage is $B=0$. Married partners incur sacrifice cost c. This cost may be, for example, temporal or permanent loss in utility from smoking.

According to Matouschek and Rasul (2008), in the short-run the gains from non-cooperative exceeds the gains from cooperation, and vice versa. Thus, each agent faces a repeated game and may play a trigger strategy. However, the agents will cooperate in any period, provided they cooperated in the previous period. If married couples cooperate at any period, the payoff to this individual would be:

$$V_m=(b-c)+\int_0^{rV_m+\gamma} rV_m dF(s)+\int_{rV_m+\gamma}^{\infty}(s-\gamma)dF(s) \tag{2.4}$$

If uncooperative married partners face no benefit or cost, then the payoff in uncooperative marriage unions would be

$$U_m = \int_0^{rU_m+\gamma} rU_m dF(s) + \int_{rU_m+\gamma}^{\infty} (s-\gamma) dF(s) \tag{2.5}$$

Thus, if the trigger strategy is an option for either agent in a marriage, expressions (2.4) and (2.5) implies that married partners will continue to cooperate if and only if:

$$b + U_m \leq V_m. \tag{2.6}$$

From equation (2.6), married couples would only cooperate if their non-deviation payoffs V_m, are greater than their deviation payoffs $b + U_m$. Equations (2.1) and (2.6) show that benefits from marriage exist, and married couples need cooperation and constant monitoring to realize the benefits.

2.2.1 Marriage and reduced propensities of smoking and drinking

Studies have shown that there are differences in morbidity and mortality rates among individuals of different marital status. Generally, among people of all race and age groups, married couples of both sexes live longer than single, divorced, or separated individuals, (Hu and Goldman, 1990; Fu and Goldman, 1996). Some studies also show that single and widowed individuals fare better than the divorced individuals in longevity (Young et al., 1998). The literature attributes better health outcomes enjoyed by married individuals to two sources: the protective effects and selection bias effects associated with marriage.

The protective effects of marriage are those benefits that married individuals enjoy because they are married. For example, married couples share risk, pool assets together and they also enjoy

social and psychological benefits associated with marriage (Fu and Goldman, 1996). Simply put, protective effects of marriage posit that being married leads to lower morbidity and mortality.

The marriage selection bias effects occur because people self-select into marriage. Healthy individuals are more likely to get married to other healthy individuals. In marriage literature, the evidence of the benefits from marriage points to lower morbidity and mortality rate among married couples, and these benefits are a combination of protective and marriage selection bias effects (Fu and Goldman, 1996; Murray, 2000; Ali and Ajilore, 2010). Controlling for the protective effects and selection bias effects of first marriage on smoking and drinking, especially at zero levels of consumption, is a central theme in my dissertation.

There is also strong evidence in the health economics literature that marriage reduces risky behaviors, as married individuals, especially women, often monitor their spouses (Umberson, 1992; Miller-Tutzauer et al., 1994; Schulenburg et al., 1995; Hu and Goldman, 1996; Leonard and Rothbard, 1999; Bachman et al., 2002; Duncan et al., 2006; Lee, 2010; Leonard, et al., 2014). Since marriages are often governed by societal social norms, married couples are expected to shun some risky behaviors as they transit from single life into married life by cooperating with each another. In most cases, marriage itself serves as an indication that an individual is willing to cooperate with another individual on several issues, including less drinking, less smoking, and less substance abuse. Thus, couples are likely to stay married if the benefits they get from cooperating with one another on issues of mutual benefits (for example, less smoking and less alcohol consumption) outweighs the short-term benefit from uncooperative behaviors.

Duncan et al. (2006) find that marriage reduces binge drinking by roughly 20% for men and 10% for women. However, the effects of cohabitation on binge drinking are mixed in their study. According to Lee et al. (2010), married couples participate less in social activities, which

in turn leads to less drinking among married couples. Empirical studies have also shown that the connection between marriage and risky behaviors like alcohol consumption and smoking is especially strong when a longitudinal dataset is employed (Williams and Umberson, 2004; Duncan et al., 2006). I use the National Longitudinal Survey of Youth 1997 (NLSY, 1997), a longitudinal dataset, in my dissertation.

In recent literature, researchers also pay attention to the impact of cohabitation among young adults on smoking and alcohol consumption. Lanardo et al. (2010) find that while cohabitation among young adults does not influence substance abuse, young adults who cohabit are likely to get less involved in delinquencies. Schulenburg et al. (1995) find that about 15-20% of adolescents who cohabitate shun chronic alcoholic abuse by the time they reach adulthood.

2.3 Past studies on modeling excess zeros

There is a rich body of analyses of tobacco and alcohol consumption in the health economics literature (Cragg, 1971; Becker and Murphy, 1988; Jones, 1989; Yen, 2005; Chaloupka and Whechsler, 1995; Aristei and Pieroni, 2007; Harris and Zhao, 2007, Aristei et al., 2008; Gurmu and Dagne, 2012). However, popular models of smoking and drinking outcomes include ordered probit or logit models. More appropriate models of tobacco and alcohol consumption include models which treat the data generating process of tobacco consumption and alcoholic beverage consumption as double hurdle models to account for the presence of excess zeros outcomes. I use the zero-inflated ordered probit (ZIOP) model, a double hurdle model, to estimate the impact of first marriage on smoking and drinking. I present the specification reasons for this choice in the subsequent chapters of this paper.

Some studies disaggregate the time profile of smoking and alcohol consumption into respondent, age and time effects (Kepteyn et al., 2005; Deaton and Paxton, 2000; Aristei et al.,

2008). Since I draw my dataset from the NLSY97, a nationally representative sample of individuals who were roughly of the same age the first time they were first interviewed in 1997, disaggregation of the age and cohort effects is not necessary. Thus, there is no need to group time profiles of smoking and drinking according to age and respondents. However, following the Becker and Murphy (1988) rational choice model, and the Harris and Zhao (2007) corner solution and double hurdle models, I use standard economic determinants of demand for goods and services, otherwise known as demand influencers. These demand influencers are own-price effect, income, and prices of related goods (substitutes and complements). I use these demand influencers as exclusion restrictions in the consumption decision equation of the ZIOP model. [3] I address these issues in the chapters 3 and 4.

2.4. Self-selection into marriage

Fu and Goldman (1996), Murray (2000), Ali, and Ajilore (2010) find that married couples live healthier life styles because individuals select into marriage. Fu and Goldman (1996) incorporate rational choice models and job search theory into their marriage selection model to show how selection bias influences marriage decisions. I model self-selection in marriage in ordered probit model and ZIOP model in the subsequent chapters of this dissertation. I also conduct tests of fit for marriage self-selection issues.

[3] The theory of rational addiction (Becker and Murphy, 1988) contends that addicts are likely to consume a health threatening quantity of the addictive good, say tobacco. Individuals in the theory of rational addiction do not also respond to a temporal change in prices but respond to permanent change in prices.

CHAPTER 3: DATA and DESCRIPTIVE STATISTICS

3.1 Summary statistics, participation, and consumption decisions with zero outcomes

The main dataset that I use in this dissertation is the National Longitudinal Survey of Youth 1997 (NLSY97). This study covers the period 1997-2013 for the drinking sample and 1997-2011 for the smoking sample. Due to lack of data on tobacco consumption in 2013, this study does not cover that period for the smoking sample. The NLSY97 survey is still ongoing, but the data for the years 2015-2017 are not currently available. Although there is limited information in the 2012 and 2014 survey (the year of marriage is one of such piece of information), the survey was not conducted in 2012 and 2014.

The NLSY97 is a nationally representative panel of 8,984 youth who were first interviewed in 1997 when they were between 12 and 16 years old. Thus, the respondents in the sample were born between 1981 and 1985. Except for the 2012 and 2014 rounds, 16 (15) rounds/waves of the NLSY97 are included in the drinking (smoking) sample between 1997 and 2013. I give the detail of how I construct the drinking and smoking samples later in this section. Thus, the NLSY97 provides a panel dataset whereby individuals are followed over time. The timeline is presented in Figure 3.1:

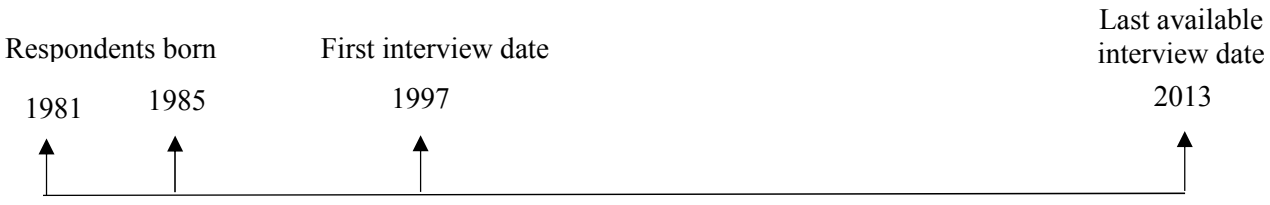

Figure 3.1: National longitudinal survey of youth

I obtained additional data for this study from other sources. The price indices of tobacco products and alcoholic beverages are obtained from the Federal Reserve Economic Data (FRED) dataset maintained by the Federal Reserve Bank in St. Louis. The data on the average street price of marijuana per gram are obtained from the Office of National Drug Control Policy (NDCP).

To observe smoking and alcohol consumption patterns in the sample, I created a time profile of five years before the first marriage year and five years after the first marriage for each married respondent in the final sample. This methodology is fully explained in Chapter 4. The data cleaning and final sample selection criteria are explained in the next four paragraphs.

A total of 1,843 (1,561) of the original 8,984 respondents have "non-interview" status as of 2013 in the drinking sample (smoking sample). This is a retention rate of 79.49% (82.6%) for the drinking sample (smoking sample) as at the 2013 (2011) survey round. According to the NLSY97 codebook, "reasons for non-interview" include death, not locatable, respondent refused to be interviewed, and incarceration. To minimize survivorship bias, all the respondents with "non-interview" status are removed from the sample.

The survey design of the NLSY97 includes dates and years of first marriage, observed zero consumption of alcohol and smoking at both participation and consumption levels, socio-economic and demographic variables, and other variables that are not interesting for this study. I apply some filters to remove the respondents with missing data for *all* the variables, both responses and predictors, used in this study. I also remove respondents with missing dates of first marriages since these dates are the reference points of the methodology of this study.

For the years 2012 and 2014, only the first marriage dates of respondents who got married in these years are available in the NLSY97. All other variables are missing for the years 2012 and 2014 in both drinking and smoking sample. For the smoking sample, tobacco consumption data is

not available for the year 2013. Since all demographic, socio-economic, smoking, and drinking response variables are not available in 2012 and 2014, I drop all respondents who married in 2012 and 2014 from this study. I also dropped the year 2013 data from the smoking sample. In total, there are 204 and 13 new first marriages in 2012 and 2014 respectively. Also, one respondent married in 1995 and another one married in 1996 before the commencement of the NLSY97. I also remove these respondents from the sample because the smoking and drinking habits as well as all the predictors used in this study are not available prior to 1997.

After removing the respondents with "non-interview" status (1,843 and 1,561 respondents from the drinking and smoking sample respectively) and respondents who experienced first marriages in the 1995, 1996, 2012 and 2014 (and 2013 for the smoking sample only), there are 6,921 respondents left in the drinking sample. In total, 3,534 (3,720) respondents in the NLSY97 have *never* experienced first marriage in the drinking sample (smoking sample) as at 2013 (2011). Based on the time profile of five years before and five years after marriage methodology, there would no reference point in terms of the years of first marriages for single respondents. So, I dropped these 3,534 (3,720) respondents from the drinking sample (smoking sample). Finally, 46 (48) respondents in the drinking sample (smoking sample) consistently refused to answer the questions about the years of their first marriages. These respondents are also removed from the final sample.

The final drinking sample (smoking sample) comprises 3,342 (3,244) respondents after the sample selection criteria described in the last four paragraphs have been implemented. Table 3.1 summarizes the procedures of each criterion. Table 3.2 presents the frequency distributions of the raw and filtered first marriage by year for the drinking sample.

Table 3.1: The selection of final sample from the NLSY97

NLSY97 Filters	Drinking	Smoking
All Respondents in the NLSY97 (1997)	8,984	8,984
Respondents with "non-interview" status	-1,843	-1,561
Respondents who married for the first time in 1995	-1	-1
Respondents who married for the first time in 1996	-1	-1
Respondents who married for the first time in 2012	-204	-204
Respondents who married for the first time in 2013	-	-168
Respondents who married for the first time in 2014	-13	-13
Respondents who have *never* been married	-3,534	-3,720
Respondents who consistently refused year of first marriage	-46	-48
Final sample	**3,342**	**3244**

In Table 3.1, the second column presents the raw distribution of first marriage by years, the third column shows the distribution of first marriage after respondents with "non-interview status" have been removed from the sample, while the fourth column shows the distribution of first marriage after respondents who have never been married and respondents who refused to answer the marriage question are removed from the sample. The years 1995, 1996, 2012, and 2014 are also removed from the fourth column of Table 3.1. I omit the smoking sample counterpart of Table 3.1 because the procedure is basically the same.

Table 3.1 also shows the subsequent revision to the years of first marriage in the NLSY97. For example, the number of respondents who married in 1995 were subsequently revised from 2 to 1 (see the first and second columns of Table 3.1). The number of refusals were also revised from 58 to 46. Figure 3.2 shows the distribution of the drinking sample of each column in Table 3.1. The density of first marriage by years in the final sample in panel C is identical to the density of the raw sample in panel A. Thus, the distributions of the drinking sample after each data management process do not change substantially from the raw sample. The smoking sample (not

Table 3.2: Distribution of First Marriage by Year (Drinking Sample)

	Raw sample	Non-interview filter	Final sample Other filters
Year of First Marriage	Frequency (%)	Frequency (%)	Frequency (%)
Non-interview/Single	4,916	3,534	-
	(54.72)	(49.49)	-
Refusal	58	46	-
	(0.646)	(0.644)	-
1995	2	1	-
	(0.0223)	(0.0140)	-
1996	1	1	-
	(0.0111)	(0.0140)	-
1997	14	13	13
	(0.156)	(0.182)	(0.389)
1998	38	31	31
	(0.423)	(0.434)	(0.928)
1999	78	65	65
	(0.868)	(0.910)	(1.945)
2000	134	116	116
	(1.492)	(1.624)	(3.471)
2001	237	203	203
	(2.638)	(2.843)	(6.074)
2002	233	202	202
	(2.593)	(2.829)	(6.044)
2003	302	261	261
	(3.362)	(3.655)	(7.810)
2004	325	278	278
	(3.618)	(3.893)	(8.318)
2005	372	328	328
	(4.141)	(4.593)	(9.814)
2006	375	333	333
	(4.174)	(4.663)	(9.964)
2007	349	304	304
	(3.885)	(4.257)	(9.096)
2008	310	282	282
	(3.451)	(3.949)	(8.438)
2009	307	277	277
	(3.417)	(3.879)	(8.288)
2010	289	246	246
	(3.217)	(3.445)	(7.361)
2011	233	211	211
	(2.593)	(2.955)	(6.314)
2012	206	204	-
	(2.293)	(2.857)	-
2013	192	192	192
	(2.137)	(2.689)	(5.745)
2014	13	13	-
	(0.145)	(0.182)	-
Total	**8,984**	**7,141**	**3,342**

shown) also follows a similar pattern. I now turn my attention to the measurements of smoking and drinking outcome responses, and the general summary statistics.

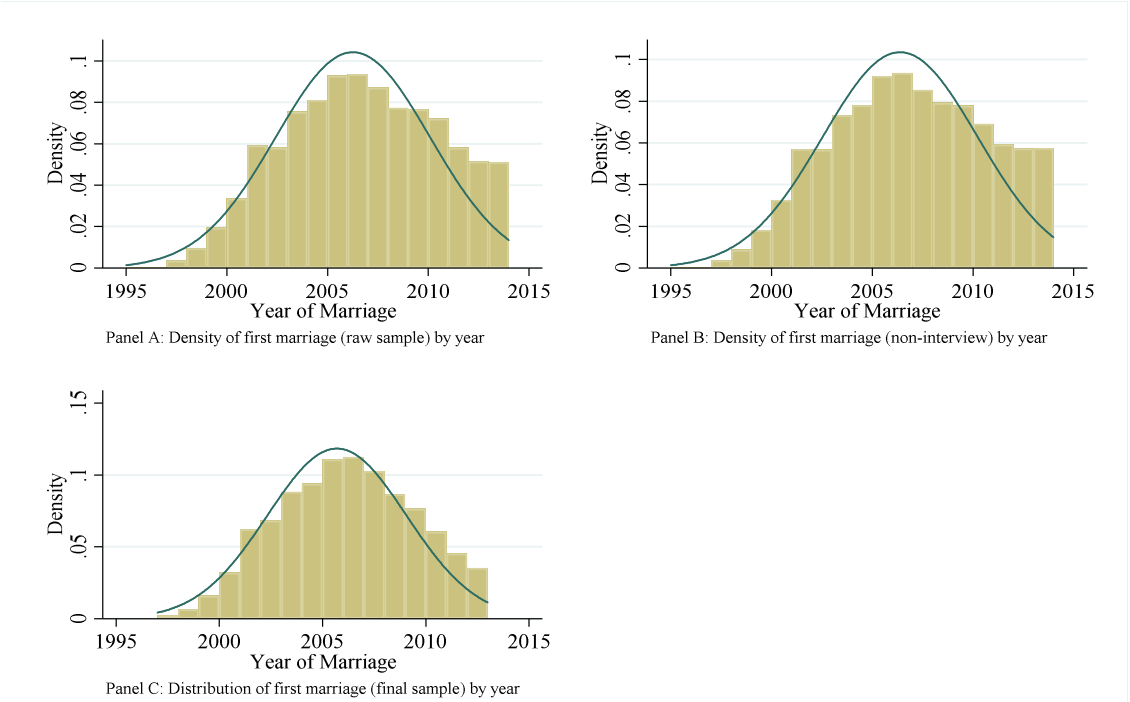

Figure 3.2: Densities of First Marriage by Year

3.2 Measurements of smoking and drinking intensities

The survey design of the NLSY97 includes questions about the consumption of alcoholic beverages and tobacco products. Although the observed zero consumption of these products are the choice response variables of this study, the NLSY97 respondents' positive consumptions are also analyzed. I build complete drinking and smoking profiles of the 3,342 (3,244) respondents from three questions on the drinking (smoking) habits of the respondents. I construct drinking intensities from three questions from the NLSY97. I build zero consumption of alcoholic beverage from two questions: "*Have you ever had a drink of an alcoholic beverage?"* and "*During the last 30 days, how many days did you have one or more drink of an alcoholic beverages?"* Respondents

who consistently answer "No" to the former question have "genuine" optimal zero consumption of alcoholic beverage. While respondents who answered "None" to the latter question are either infrequent/occasional drinkers or drinkers with corner solution optimal consumption. Positive amounts of alcoholic beverage consumed by the respondents are constructed from the question "*In the past 30 days, on the days when you drink alcohol, about how many drinks did you usually have?* I then convert the raw smoking and drinking outcomes to ordinal outcomes using the methodology described in the next couple of paragraphs.

Similarly, for smoking intensities, I build zero consumption from two questions: "*Have you ever smoked a cigarette?"* and "*During the last 30 days, how many days did you smoke a cigarette?"*. Respondents who consistently answer "No" to the former question have perfectly inelastic demand for cigarettes, and for these respondents zero demands for cigarettes are optimal choices. This is the first type of zeros ("abstention zeros") that I identified in chapter one. Respondents who answer "None" to the latter question belong to one of these two categories: those whose optimal choice of zero consumption are defined by corner solutions and infrequent/occasional smokers. I then construct positive amounts alcoholic beverage consumption from the question "*When you smoked during the last 30 days, how many cigarettes did you usually smoke each day?"*

Let Y_{it}^{d} and Y_{it}^{S} be measures of alcoholic beverage and tobacco consumptions, respectively, by an individual *i* at time *t*. Thus, Y_{it}^{d} is the number of drinks of alcoholic beverage consumed on the days that individual *i* drinks, including zero consumption. Y_{it}^{S} is number of sticks of cigarettes smoked on days that individual *i* smokes, including zero consumption. For example, in the drinking sample, $i = 1,2,3, \dots, 3{,}342$ and $t = 1,2, \dots, 11$. Time *t* is a profile that I created around the year of first marriage. The median year, *t=6*, is the year of marriage for an individual *i* in the dataset,

$t \in [1,5]$ is the 5-year period before the first marriage of this individual, and $t \in [7,11]$ is the 5-year period after the first marriage of this individual. I describe these time profiles later in this section and over the next two chapters. Next, I convert the raw Y_{it}^{d} and Y_{it}^{S} into ordinal drinking and smoking intensities labeled 0, 1, 2, 3.

Reporting the frequency distribution of the raw Y_{it}^{d} and Y_{it}^{S} measurements does not really help in summarizing the smoking and drinking intensities for empirical analysis as some individuals in the NLSY97 reportedly took 99 drinks of alcoholic beverages on the days they drink. Some respondents also claimed that they smoked 99 sticks of cigarette on the days they smoke. More so, converting raw smoking and drinking numbers to 0-3 ordinal one-unit interval is common in smoking literature (see Harris and Zhao, 2007; Aritistei et.al, 2008). Thus, I construct four ordinal outcomes (0-3) of drinking and smoking intensities from their respective observed quantities, Y_{it}^{S} and Y_{it}^{d}.

For drinking intensities, I construct the dependent ordinal variable of drinking intensities Y_{it}^{0d} with the values $Y_{it}^{0d} = 0$ if an individual is not a current drinker or has never drank before, $Y_{it}^{0d} = 1$ if an individual drinks at least once in a week (7 days), $Y_{it}^{0d} = 2$ if an individual has 5 or more drinks more once in a week, but less than daily binge drinking, and $Y_{it}^{0d} = 3$ if the individual drinks daily with more than 5 drinks.[4] The ordinal drinking intensities are summarized in equation (3.1):

$$Y_{it}^{0d} = \begin{cases} 0 & \text{if indvidual } i \text{ is not a current drinker or has never drank} \\ 1 & \text{if individual } i \text{ drinks weekly or less} \\ 2 & \text{if individual } i \text{ drinks daily, and } Y_{it}^{d} < 5 \\ 3 & \text{if individual } i \text{ drinks daily, with } Y_{it}^{d} \geq 5 \end{cases} \tag{3.1}$$

[4] The Centre for Disease Control (CDC) uses different measures of binge drinking for men and women. For men, the CDC defines binge drinking as having 5 or more drinks in about 2 hours. For women, it is 4 or more drinks in 2 hours according to CDC. However, I do not make such distinctions in this paper.

For smoking intensities, I construct a dependent ordinal variable of smoking intensity Y_{it}^{0S} with the values $Y_{it}^{0S} = 0$ if the individual is not a current smoker or has never smoked before, $Y_{it}^{0S} = 1$ if the individual smokes weekly (7 days) or less, $Y_{it}^{0S} = 2$ if the individual smokes daily with less than 20 sticks of cigarette (1 pack of cigarette), and $Y_{it}^{0S} = 3$ if the individual smokes daily with more than 20 sticks of cigarette. The ordinal smoking intensities are summarized in equation (1):

$$Y_{it}^{0S} = \begin{cases} 0 & \text{if indvidual } i \text{ is not a current smoker or has never smoked} \\ 1 & \text{if individual } i \text{ smokes weekly or less} \\ 2 & \text{if indivudual } i \text{ smokes daily, and } Y_{it}^{S} < 20 \\ 3 & \text{if an individual } i \text{ smokes daily, with } Y_{it}^{S} \geq 20. \end{cases} \quad (3.2)$$

The ordinal distribution of drinking and smoking intensities summarized in equations (3.1) and (3.2) are shown for selected years 1997, 2000, 2002, 2010, and 2013 from the drinking sample, and (1997, 2000, 2002, 2010, and 2011) from the smoking sample in Tables 3.4 and 3.3 respectively. T

The proportion of missing values are quite high in the early years in Tables 3.3 and 3.4 because several minors were exempted from the drinking and smoking questions in the early rounds of the NLSY97 due to legal and privacy concerns. But these missing values gradually reduced over the years as these minors attain adulthood in later rounds of the NLSY97. Overall, the proportion of respondents who did not smoke in the last 30 days rose as the years go by, while the converse is true for the proportion of drinkers who did not drink in the last 30 days. That is, it seems that the respondents are more likely to be infrequent smokers than infrequent drinkers as the years went by. Also, the proportion of respondents who smoke weekly or less falls over the years, while the opposite is true for respondents who drinks weekly or less. It also apparent from Table 3.3 that there is high proportion of 'nonparticipating' zeros in the smoking sample over the years. But the proportion of 'nonparticipating' zero drinkers fall over the years in Table 3.4.

To get a clearer picture of the overall trend in smoking and drinking intensities, I plot the 'smoothened' raw average of smoking and drinking measurements, Y_{it}^{S} and Y_{it}^{d}, against first marriage event window timeline in Figures 3.3 and 3.4. To 'smooth' the plot of the raw amount of cigarette consumed (raw amount of alcoholic beverage consumed), I restricted Y_{it}^{S} (Y_{it}^{d}) to twenty (five), the number of cigarettes in a pack (the binge drinking threshold).

Figure 3.3: Stick of Cigarette Smoked

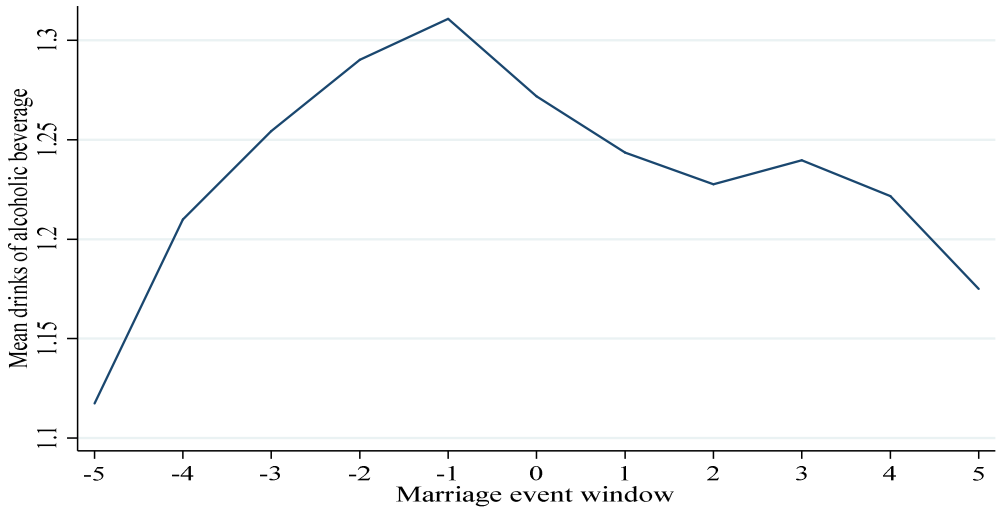

Figure 3.4: Drinks of Alcohol Beverage < 5

Table 3.3: Distribution of Smoking Intensities

Drinking Intensities	Ordinal Outcome	1997		1999		2002		2010		2011	
		N	%	N	%	N	%	N	%	N	%
Any smoke in the last 30 days? (1=No)	0	288	44.72	633	56.37	1,144	56.44	1,351	68.75	1,162	69.08
Days smoked in the last 30 days (1=None)	0	171	26.55	108	9.62	133	6.56	91	4.63	83	4.93
How many cigars if respondent smokes weekly or less	1	110	17.08	179	15.94	364	17.96	229	11.65	202	12.01
How many cigars if respondent smokes weekly, < 20	2	75	11.65	199	17.72	372	18.35	286	14.55	227	13.50
How many cigars if respondent smokes weekly, >20	3	.	.	4	0.36	14	0.69	8	0.41	8	0.48
Subtotal		**644**		**1,123**		**2,027**		**1,965**		**1,682**	
Missing values		2,600		2,121		1,217		1,279		1,562	
Never married		3,744		3,744		3,744		3,744		3,744	
Refusals (First marriage question)		48		48		48		48		48	
Non-interview		1,561		1,561		1,561		1,561		1,561	
Married in 1995, 1996, 2012, 2013, 2014		387		387		387		387		387	
Total		**8,984**		**8,984**		**8,984**		**8,984**		**8,984**	

Table 3.4: Distribution of Drinking Intensities

Drinking Intensities	Ordinal Outcome	1997		1999		2002		2010		2013	
		N	%	N	%	N	%	N	%	N	%
Any drink in the last 30 days? (1=No)	0	276	43.95	441	41.29	588	30.11	460	22.72	296	20.10
Days drunk in the last 30 days (1=None)	0	190	30.25	171	16.01	253	12.95	194	9.58	136	9.23
How many drinks if respondent drinks weekly or less	1	117	18.63	343	32.12	855	43.78	1,205	59.51	950	64.49
How many drinks if respondent drinks > weekly, < 5 drinks	2	42	6.69	111	10.39	248	12.70	165	8.15	89	6.04
How many drinks if respondent drinks > weekly, > 5 drinks	3	3	0.48	2	0.19	9	0.46	1	0.05	2	0.14
Subtotal		**628**		**1,068**		**1,953**		**2,025**		**1,473**	
Missing values		2,714		2,274		1,389		1,317		1,869	
Never married		3,534		3,534		3,534		3,534		3,534	
Refusals (First marriage question)		46		46		46		46		46	
Non-interview		1,843		1,843		1,843		1,843		1,843	
Married in 1995, 1996, 2012, 2014		219		219		219		219		219	
Total		**8,984**		**8,984**		**8,984**		**8,984**		**8,984**	

The marriage timeline window on the horizontal axes of Figures 3.3 and 3.4 are different from the ones that I briefly described in the third paragraph of this section. Here, the first marriage event window timeline of 0 is equivalent to the year of first marriage (t=6 in the third paragraph) above, the pre- marriage event window timelines are the years before first marriage ($t \in [1,5]$) above, while the post marriage event window timelines are the years after first marriage ($t \in [7,11]$). A cursory glance at Figures 3.3 and 3.4 reveals three distinct trends: a rise in smoking and drinking before the year of first marriage, a fall in smoking and drinking around the year of first marriage, and another steep fall in smoking and drinking intensities after the year first marriage. The methodology, hypotheses and the baseline regression of this study are all built on these three trends.2

3.3 Summary statistics

I present the summary statistics of smoking and drinking response choice variables and all their predictors in this section. The number of respondent-year observations reported for these variables depend on three factors. First, the number of respondents in the final drinking sample (3,342 respondents) and smoking sample (3,244 respondents). Second, the 10-year time profile is created around the year of first marriage. Third, there are missing values in the smoking response outcomes, in the drinking response outcomes, and in the predictor variables. If there were no missing variables, and all the respondents in the final sample experience first marriages between the years 2002 (5 years after 1997) and 2007 (5 years before 2011), there would be a maximum of 36,762 (3342×11) and 35,684 (3244×11) respondent-year observations for the drinking and smoking samples respectively. However, the total respondent-year observations are less than these maximum possible because some respondents have missing values. Besides, all the respondents in

the sample are included in the baseline regression irrespective of how long they have been married. In other words, the sample panel of this study is not balanced.

Since univariate analysis is not sufficient to establish causality, later chapters incorporate the econometric analysis of the effects of demographic and economic variables on smoking and drinking response variables. But Table 3.5 shows the summary statistics of demographic and economic variables of the drinking sample, including age, education, wages, and price indices of tobacco and liquor drinks. There is no need to show separate summary statistic tables for the drinking and smoking samples because they both have identical predictors and baseline regressions. The average age and wages in the sample are 23.58 years (24.01 years for men, 23.22 years for women) and $15,808.00 ($20,331.60 for men, $12,100.72 for women) respectively. Forty-five (45%) of the final sample is male. The average age in the sample appears to be low because the summary statistics is calculated from a pooled data. The average age in the sample is 31.06 years if one computes the age variable summary statistics for only the year 2013.

The racial composition of the final sample is 60%, 17%, 22% and 1% for whites, blacks, Hispanics and people of mixed races respectively. Other demographic and socioeconomic variables are mostly qualitative in nature, and these variables are defined alongside their qualitative measures in the first column of Table 3.5. Heights and weights of the respondents are also presented in Table 3.5. Heights and weights are included as part of exclusion restrictions variables in the marriage selection equation. Following Mayer et al. (2010), I added a dummy variable for health insurance to control for moral hazard. Other variables used for this study include a dummy for employment status, a dummy for self-reported health status, the number of under 6 years old children in the household, consumer price indices of tobacco and alcoholic beverages, and the street price of a kilogram of marijuana.

Table 3.4. Summary Statistics

variables	N	mean	SD	Min	Max
Ordinal Drinking Variable (All Sample)	29,908	0.719	0.648	0	3
Ordinal Zero Alcohol Consumption	11,556	0	0	0	0
Ordinal Positive Alcohol Consumption	18,352	1.172	0.392	1	3
Ordinal Smoking Variable (All Sample)	28,749	0.495	0.784	0	3
Ordinal Zero Tobacco Consumption	19,588	0	0	0	0
Ordinal Positive Tobacco Consumption	9,161	1.553	0.530	1	3
Married Year	32,565	2006	3.371	1997	2013
Gender (Male=1, Female=2)	32,565	1.550	0.498	1	2
Male (=1)	14,666	0.450	0.498	0	1
Race (All sample)	32,565	3.044	1.220	1	4
Black (=1)	5,419	0.166	1.222	0	1
Hispanic (=1)	7,286	0.224	1.222	0	1
Mixed Race (=1)	303	0.009	1.222	0	1
White (=1)	19,557	0.601	1.222	0	1
Age (in Years)	32,565	23.58	3.978	13	33
Education (All sample)	29,612	1.103	0.694	0	3
Education - High School or Less (=1)	22,105	0.774	0.418	0	1
Education - Bachelor's Degree (=1)	5,533	0.196	0.418	0	1
Education – Greater than Bachelor's (=1)	638	0.024	0.418	0	1
Marital Status	29,241	2.508	0.943	1	6
Single, Cohabitation (=1)	3,769	0.129	0.943	0	1
Single, Non-cohabitation (=1)	10,042	0.343	0.943	0	1
Married (=1)	13,847	0.474	0.943	0	1
Separated (=1)	784	0.027	0.943	0	1
Widowed (=1)	799	0.027	0.943	0	1
Cohabit (Yes=1)	32,565	0.130	0.336	0	1
Employment Status	31,784	2.675	0.595	1	3
Unemployed (=1)	2,135	0.067	0.595	0	1
Out of the Labor Force (=1)	6,065	0.191	0.595	0	1
Employed (=1)	23,584	0.742	0.595	0	1
How Many Under 6 years kid in Household?	30,345	0.564	0.842	0	8
Any Insurance? (1=Yes)	32,565	0.571	0.495	0	1
Self-Reported Health Status (1=lowest)	30,387	2.124	0.926	1	5
Wage ($)	32,565	15,808	21,190	0	180,331
Height (in Feet)	28,587	5.164	0.412	2	7
Weight (in Pounds)	32,565	147.8	70.83	0	999
CPI - Alcoholic Beverage	32,565	200.0	19.19	162.8	234.6
CPI – Tobacco Products	32,565	556.6	159.4	243.7	876.8
Average Street Price of Marijuana ($)	32,565	16.08	1.598	12.35	19.35

CHAPTER 4: METHODLOGY

In this chapter, I describe the methodology, equations and hypotheses of this study. To estimate the impact of first marriage on tobacco and alcohol beverage consumption, I create a time profile of ten years around the year of first marriages of the respondents in the NLSY97. This time profile is built on five years before the year of first marriage and five years after the year of first marriage. I begin this chapter with detailed explanations of these time profiles. I then identify four hypotheses on how the time profiles are linked to one another. I also examine the effects of these timelines on the probabilities of reduced smoking and drinking. After setting up these hypotheses, I specify and describe the ordered probit (OP) model, the ordered probit with endogenous dummy (OP-ED) model, and the zero-inflated ordered probit (ZIOP) model. I describe the sample selection ordered probit (SSOP) model in Chapter 5.

4.1 Time profiles around the year first marriage

The NLSY97 respondents are followed for a period of eleven years namely: five years before the year of first marriage, the year of the first marriage, and five years after the year of first marriage. Thus, $t \in [1,11]$, where t is the time. The median year is the year of first marriage, that is $t = 6$ is the year of first marriage. Years $t \in [1,5]$ are the years before the year of first marriage, while the years $t \in [7,11]$ are the years after the year of first marriage. Here, smoking and drinking response variables are isolated within a 10-year interval around the year first marriage. Creating these time profiles made it easier for me to isolate drinking and smoking outcomes into pre-and post-first marriage periods. For example, if an individual is married for the first time in 2005, his smoking and drinking habits in 2000 are recorded as occurring in year $t = 1$, while his drinking

and smoking habits in 2009 and 2010 are recorded as occurring in years $t = 10$ and $t = 11$ respectively. This methodology is similar to the one employed by Duncan et al. (2006).

However, unlike Duncan et al. (2006), I use the double hurdle models of participation and consumption decisions with emphasis on zero consumption in a similar manner as the Harris and Zhao's ZIOP model. But unlike Harris and Zhao (2007), I use a spline regression function as the baseline regression. I also address the issue of endogenous right-hand side variable within the OP model. Moreover, I also examine the SSOP model for sample attrition and survivorship bias in Chapter 5. Before specifying these models, I build a theory about the nature of the trend of the time profiles around the year of first marriage, and I use a linear spline function to specify the trends around the year of first marriage.

A linear spline function is a continuous function with different linear slopes at different segments of the function. Each of these linear segments are connected at a point called knots. Using a spline function, I join the theoretical slopes of first marriage time profiles into a single continuous function shown in Figure 4.1 below. A casual glance at this figure shows a spline function of drinking and smoking behaviors around the year of first marriage. In all, there are three different slopes on the spline function: the slope before first marriage, the slope around first marriage, and the slope after first marriage.

The spline function that I adopt in this study is an attempt at mimicking Figures 3.3 and 3.4 of Chapter 3. The theoretical predictions of the relative magnitudes of these different slopes of the spine function are part of the statements of hypotheses of this study. I specify the equation of the spline function in Figure 4.1, and this specification is the baseline regression of this study. This baseline regression can be found in section 4.2.1. Derivations of the spline functions for the OP model, the ZIOP model, the OP-ED model, and the OP-SS model are also shown in section 4.2.1.

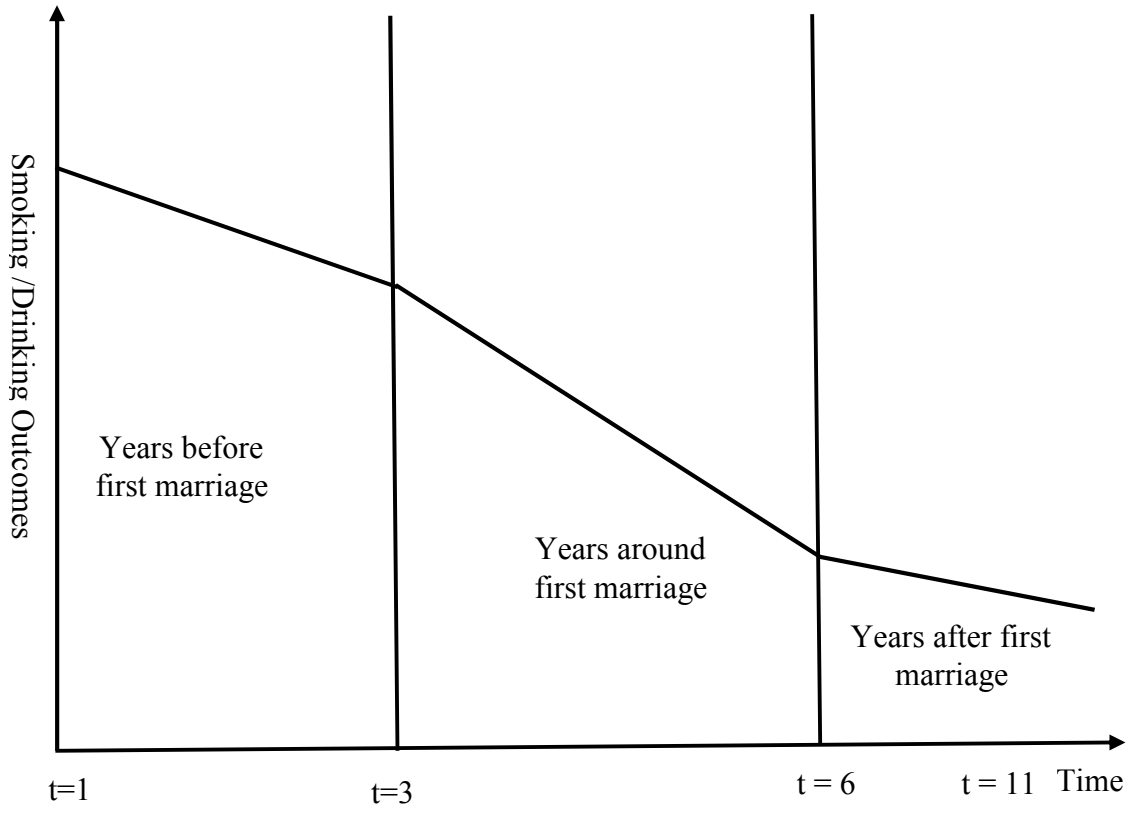

Figure 4.1: A Spline Function of Smoking and Drinking Outcomes Around First Marriage

In Figure 4.1 above, married individuals sharply reduce their smoking and drinking habits around their years of first marriages. The "years around first marriage" is a concept that I loosely define to capture the drinking and smoking habits of the respondents between the middle year before first marriage and the year of first marriage. That is, the "years around first marriage" is the period $t \in [3,6]$, or three years preceding the year of first marriage. The years before first marriage is the period $t \in [1,2]$. The years after first marriage is $t \in [7,11]$, or 5 years after the year of first marriage. The choice of the three years preceding the year of first marriage as the boundary year

between the "years before first marriage" and the "years around first marriage" is the median year of the years before first marriage, i.e. $t \in [1,5]$.

From the foregoing analysis, as the respondents get closer to their years of first marriage, they begin to 'clean up their acts' by reducing their drinking and smoking habits. These smoking and drinking trends continue after the year of first marriage. The relative size of the marriage effect on smoking and drinking behavior around and after first marriage is a matter of hypothetical deduction. I attempt to capture these deductions by the statements of hypotheses later in this section.

An alternative method to the spline function specification is splitting the sample into two (before and after first marriage) and running a separate regression for each subsample. But this method will achieve little in helping one understand the nature and magnitude of the different slopes and continuity of Figure 4.1. Also, splitting the sample would not capture the behavioral patterns of smoking and drinking outcomes around the first year of marriage. It is also tempting to specify a difference in difference (D-in-D) estimation method to estimate the impact of smoking and drinking on marriage and smoking. Like the sample splitting method, this method will achieve very little in joining the knots at the end of each slope of the spline function in Figure 4.1.

To capture the slopes of the theoretical spline function illustrated Figure 4.1, I use the interaction of time $t \in [1,11]$ and first marriage dummy variable, and I draw the first two set of hypotheses based on spline function in Figure 4.1. I draw the third and fourth hypotheses from the literature review in Chapter 2.

Hypothesis 1(H1): The probability of zero tobacco consumption (smoking) does not increase around and after the year of first marriage.

Hypothesis 2 (H2): The magnitude of probability of zero tobacco consumption (smoking) around and after first marriage is not greater for males than females.

Hypothesis 3 (H3): The probability of zero alcoholic beverage consumption (drinking) does not increase around and after the year of first marriage.
Hypothesis 4 (H4): The magnitude of the estimated probability of zero alcoholic beverage consumption (drinking) around and after first marriage is not greater for men than women.

The last two hypotheses simply imply that males benefit more than females in terms of reduction in drinking and smoking around and after the year of first marriage. The last two hypotheses are drawn from the literature of risky behavior and marriage in Chapter 2 (See Duncan (2006), Ali and Ajilore (2010), Murray (2000), Leonard et al. (2014)). The theoretical derivations from the work of Matouschek and Rasul (2008) summarized in Chapter 2 also supports these claims.

4.2 Drinking and Smoking outcomes, and their baseline regression models

The baseline regression of this study is based on the theoretical spline function that I introduced in Figure 4.1. I specify the baseline regression for the ordered probit (OP), the ordered probit endogenous dummy (OP-ED), the zero-inflated ordered probit (ZIOP) models in this section.

Since there are three slopes in the segments on the theoretical spline function, I need $n - 1 = 2$ knots to maintain the continuity of the theoretical spline function in Figure 4.1. While the positions of the knots can be chosen as free parameters, one can also treat the position of these knots as choice variables (see Wold, 1974). Since the knots can be chosen as choices variables, I use the sample data for the location of the knots. I use the median of the years before the year of first marriage and after first marriage as the location of two the knots on the spline function. These medians are also the means values of the first marriage variable. These knots correspond to $t_1^* = 3$ and $t_2^* = 6$ on the theoretical spline function in Figure 4.1.

Separate baseline specifications are required to model the impact of first marriage on smoking and drinking. But I only derive the specification of the smoking baseline regression. One can generalize the baseline smoking model to the baseline drinking model by substituting the drinking response variable Y_{it}^{0d*} for the smoking response variable Y_{it}^{0S*}. Let Y_{it}^{0S*} be the continuous latent smoking intensity. The spline specification of the impact of first marriage on smoking is:

$$\begin{aligned} Y_{it}^{0S*} &= \beta_0{}^0 + \beta_1{}^0 t + \underline{\pi^0}\,\underline{X_{it}} + \varepsilon_{it,Y^{0s}}^0 && \text{if } t < 3 \\ Y_{it}^{0S*} &= \beta_0{}^1 + \beta_1{}^1 t + \underline{\pi^1}\,\underline{X_{it}} + \varepsilon_{it,Y^{0s}}^1 && \text{if } 3 \le t < 7 \\ Y_{it}^{0S*} &= \beta_0{}^2 + \beta_1{}^2 t + \underline{\pi^2}\,\underline{X_{it}} + \varepsilon_{it,Y^{0s}}^2 && \text{if } t \ge 7. \end{aligned} \tag{4.1}$$

Since $t \in [1,5]$ is the period before first marriage, $t = 6$ is the year of first marriage and $t \in [7,11]$ is the period after marriage, a marriage dummy variable can be created from the time variable t. The intercepts or 'jumps' at each knot in Figure 4.1 are constructed from the time variable, *t*. $\underline{X_{it}}$ is a covariate vector of socioeconomic and demographic variables that affect smoking intensity. The components of $\underline{X_{it}}$ variables are shown in the summary statistics in Table 3.5 of Chapter 3. $\varepsilon_{it,Y}$ is the unobserved error term. Before I specify the baseline regression of this study from equation 1, I need to mention two points about the nature and derivation of the theoretical and estimated slopes of the spline function in Figure 4.1.

To derive the estimates of the theoretical slopes and intercepts in Figure 4.1, I specify a spline function of equation (4.1) above. That is, I break up the marriage dummy variable into three, one for each of the three slopes in Figure 4.1. At the threshold values or knots of $t_1^* = 3$ and $t_2^* = 6$, let $d_1 = 1$ if $t < t_1^*$, $d_2 = 1$ if $t_1^* < t \le t_2^*$, and $d_3 = 1$ if $t \ge t_2^*$. Thus, the coefficients of the dummy variables d_1 , d_2, and d_3 are, respectively, the intercepts or 'jumps" of Figure 4.1 during the pre-first marriage period, around first marriage, and after first marriage period. The estimates

of the slopes of the spline function in equation (4.1) are the interactions of the spline dummy variables and time. These interaction terms are the impact of time on the probability of smoking in cases involving married smokers before, around, and after their first marriages. These interaction terms occur at the two knots. Combining all the three equations in equation (4.2):

$$Y_{it}^{0S*} = \alpha_0 + \alpha_1 t + \gamma_2 d_2 + \theta_2 d_2 t + \gamma_3 d_3 + \theta_3 d_3 t + \underline{\pi}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}} \ . \quad (4.2)$$

From equation (4.2), the slopes of the three segments at the knots are: α_1, $\alpha_1 + \theta_2$, and $\alpha_1 + \theta_2 + \theta_3$. The 'jumps' or intercepts on Figure 4.1 occur at α_0, $\alpha_0 + \gamma_2$, and $\alpha_0 + \gamma_2 + \gamma_3$. It should be noted that d_1 is omitted from equation (4.2) to fulfil the full rank condition of the baseline regression in equation (4.2).

Joining the spline function in equation (4.2) at the knots results in equations (4.3) and (4.3A) below:

$$Y_{it}^{0S*} = \ \alpha_0 + \alpha_1 t + \theta_2 d_2 (t - t_1^*) + \theta_3 d_3 (t - t_2^*) + + \underline{\pi}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}. \quad (4.3)$$

$$Y_{it}^{0S*} = \alpha_0 + \alpha_1 t + \theta_2 d_2 (t - 3) + \theta_3 d_3 (t - 6) + \underline{\pi}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}. \quad (4.3A)$$

The full derivation of equation (4.3) is shown in appendix A. Equation (4.3) is the baseline regression. The constraints $t - 3$ if $t \geq 3$ and 0 otherwise, and $t - 6$ if $t \geq 6$ and 0 otherwise apply. After plugging the constraints into equation (4.3) gives equation (4.3A), the coefficient of interest in equation (4.3A) are α_0, α_1, θ_2, and θ_3. α_1 is the slope when $t < 3$. θ_2 is the change in the slope when an individual enters the "years around first marriage" period. That is, θ_2 is the relative difference in the probability of smoking for an individual during "years before first marriage" and "around the year of first marriage". Thus, $\alpha_1 + \theta_2$ is the slope of "years around first marriage" segment of the spline function in Figure 4.1 and equation (4.3A). θ_3 can be interpreted in a similar manner.

The derivation of the baseline regression of drinking intensities follows the same steps as equations (4.1) to (4.2). Let Y_{it}^{0d*} be the continuous latent drinking intensity. After plugging the latent drinking intensity, Y_{it}^{0d}, in equations (4.3) and (4.3A), the baseline regression of the impact of first marriage on drinking is:

$$Y_{it}^{0d*} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^*) + \theta_3 d_3(t - t_2^*) + + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0d}}. \quad (4.4)$$

$$Y_{it}^{0d*} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - 3) + \theta_3 d_3(t - 6) + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0d}}. \quad (4.4A)$$

All the variables in equations (4.4) and (4.4A) are just as defined in equations (4.1) to (4.3).

4.3 The OP and OP-ED model specifications of smoking and drinking intensities

I present the baseline specifications of the OP and OP-ED models in this section. I build these models on the baseline regressions shown in the last section. While the OP model is the model specification for abstention zeros identified in Chapter 1, OP-ED model specifies the possibility of self-selection in first marriages. To determine which of these models is the correct specification, I run a horserace between the OP and OP-ED models in Chapter 6. I include the four demand-influencing variables that I analyzed in Chapter 1 in the OP and OP-ED models. These variables are the Consumer Price Index (CPI) of cigarette and tobacco products, income, the price of complements (the street price of marijuana), and the price of substitutes (the CPI of alcoholic beverage). Just as in section 4.1, I only give full description of the OP and OP-ED analysis for smoking response specifications in this section. I then generalize the model specifications to drinking outcomes.

4.3.1 The OP model specification

Let Y_{it}^{0S*} be the continuous latent ordinal outcome of smoking, Y_{it}^{S}. Y_{it}^{0S*} is connected to Y_{it}^{S} via:

$$Y_{it}^{0S} = \begin{cases} 0 & \text{if } Y_{it}^{0S*} \leq 0 \\ 1 & \text{if } 0 < Y_{it}^{0S*} \leq \mu_1 \\ 2 & \text{if } \mu_1 < Y_{it}^{0S*} \leq \mu_2 \\ 3 & \text{if } \mu_2 < Y_{it}^{0S*}. \end{cases} \quad (4.5)$$

Y_{it}^{S} has been described in section 3.2 as the raw and unordered measurement of smoking intensities. I transform these raw measures of smoking intensities in equation (3.1) of section 3.2 in Chapter 3. Hence, the smoking response variable takes ordinal outcomes $Y_{itj}^{0S}, j = 0,1,2,3$. These ordinal values are described in equation (4.5):

$$Y_{it}^{0S} = \begin{cases} 0 & \text{if indvidual } i \text{ is not a current smoker or has never smoked} \\ 1 & \text{if individual } i \text{ smokes weekly or less} \\ 2 & \text{if indivudual } i \text{ smokes daily, and } Y_{it}^{0S*} < 20 \\ 3 & \text{if an individual } i \text{ smokes daily, with } Y_{it}^{0S*} \geq 20. \end{cases} \quad (4.6)$$

For an OP model, the ordinal values of the observed smoking intensities, Y_{it}^{0S}, is determined by the thresholds shown in equation (4.5) above. $\mu_0, \mu_1, \mu_2, \mu_3$ are constants (thresholds) which are also estimated in the model. The standard OP model assumptions about the thresholds hold. That is μ_{-1}=$-\infty$, μ_0=0, and μ_J=+∞. $\mu_1 < \mu_2$ also apply to equations (4.3) and (4.4). All the variables in equations (4.5) and (4.6) are as previously described in section 4.1.

$$Y_{it}^{0S*} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^*) + \theta_3 d_3(t - t_2^*) + + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0S}}. \quad (4.7).$$

Equations (4.6) and (4.7) is a standard OP model of smoking intensity. All the parameters and variables in equation (4.7) are as previously described in section 4.1. To model the abstention zeros, I run the regression of the OP model in equation (4.7), including all the four variables economic variables that I identified in Chapter 1 (CPI of tobacco, income, CPI of alcoholic beverage (a complement), and the street price of marijuana (a substitute)). The economic and theoretical explanations for the inclusion of these variables can be found in Chapter 1.

One can do a joint test of the four exclusion restrictions: $\pi_{CPI\_Tobacco} = 0, \pi_{CPI\_Alcohol} = 0, \pi_{Price\_marijuana} = 0$, and $\pi_{income} = 0$. This joint test is conducted in Chapter 5. A rejection of the null hypothesis $H_0: \pi_{CPI\_Tobacco} = \pi_{CPI\_Alcohol} = \pi_{Price\_marijuana} = \pi_{income} = 0$ will imply that at least one of the four standard theory of demand variable does not impact the smoking intensities.

Let Y_{it}^{0d*} be the continuous latent ordinal outcome of drinking, Y_{it}^{d}. Y_{it}^{d} has been described in section 3.2 as the raw and unordered measurement of drinking intensities. I transform these raw measures of smoking intensities in equation (3.2) of section 3.2 of Chapter 3. Hence, the drinking response variable takes ordinal outcomes $Y_{itj}^{0d}, j = 0,1,2,3$. These ordinal values are described in equation (8) below:

$$Y_{it}^{0d} = \begin{cases} 0 & \text{if indvidual } i \text{ is not a current drinker or has never drinked} \\ 1 & \text{if individual } i \text{ drinks weekly or less} \\ 2 & \text{if individual } i \text{ drinks daily, and } Y_{it}^{0d*} < 5 \\ 3 & \text{if individual } i \text{ drinks daily, with } Y_{it}^{0d*} > 5. \end{cases} \quad (4.8)$$

Thus, we have the baseline of the OP model from equations (4.8) and (4.9) below:

$$Y_{it}^{0d*} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^*) + \theta_3 d_3(t - t_2^*) + + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0d}}. \quad (4.9)$$

4.3.2 The OP-ED model and self-selection in first marriage search process

I present the derivation of the baseline equation of the ordered probit with endogenous dummy (OP-ED) model in this section. The OP model cannot adequately fit the impact of first marriage on drinking and smoking if the marriage search process is not random. This section describes the OP-ED model, an alternative model to the OP model if married individuals self-select into marriage. That is, in this section I consider the possibility of inconsistent OP model estimates if individuals self-select and marry people of the same background, similar intelligence,

similar physical attributes, etc. Just like the OP model in section 4.2.1, I build the OP-ED model on the baseline regressions of smoking and drinking intensities of section 4.1.

To test the fit of the OP-ED model, one needs to examine the error term sphericity assumption in the OP equations (4.7) and (4.9) of section 4.2.1. $\varepsilon_{it,Y^{os}}$ of equation (4.7) is assumed to be independent and identically distributed across respondents and time in the standard OP models that are specified in equations (4.7) and (4.9) of section 4.2.1. This assumption implies that there is no effect across time for each respondent and there is no 'within effect' in the sample of the NLSY97. Thus, pooled regression is assumed to be a good fit for the NLSY panel data set. However, this assumption is somewhat difficult to justify due to the panel nature of the sample of this study. But this assumption is frequently used for double hurdle and maximum simulated likelihood (MSL) models (See Harris and Zhao, 2007; Bratti and Miranda, 2011; Roodman, 2011).

Let M_{it} be a binary variable equal to 1 if $t \in [6,11]$, and zero otherwise. That is, $M_{it} = 1$ if a respondent has experienced first marriage, and 0 otherwise. If the marriage search process is not random, the first marriage dummy, M_{it}, is not exogenous in equations (4.7) and (4.9) of section 4.2.1. Consequently, $E[\varepsilon_{it,Y}|M_{it}, d_1, d_2, t, X_{it}] = 0$ will not hold, and the estimates of equations (4.7) and (4.9) will be inconsistent if the marriage search process is not random. In this case, a separate self-selection equation of marriage is jointly estimated with the OP models in equation (4.7). The marriage self-selection equation is generated by a latent dummy, M_{it}^* via equation (10) below:

$$M_{it}^* = \delta W_{it} + \varepsilon_{it,M}. \tag{4.10}$$

where $M_{it} = \begin{cases} 1 & \text{if } M_{it}^* > 0 \\ 0 & \text{otherwise} \end{cases}$.

The joint estimation of equations (4.7) and (4.10) is tantamount to estimating an OP model with an endogenous dummy variable. This model is the ordered probit dummy variable (OP-ED)

model. Using similar arguments, I extend the OP-ED specification in equation (4.10) to the drinking intensities equation. M_{it} is as previously defined in equation (7). $\varepsilon_{it,M}$ is the error term of the marriage latent outcome in equation (10), with $\varepsilon_{it,M} \sim N(0,1)$. $\underline{W_{it}} = [\underline{X_{it}}, \underline{Z_{it}^{S0}}]$ is a covariate of socioeconomic and demographic variables that affects the probability of first marriage event. $\underline{W_{it}}$ includes all the variables in vector $\underline{X_{it}}$ of equations (4.7) and (4.9). W_{it} also include the variables in the exclusion restriction vector $\underline{Z_{it}^{S0}}$.

For identification of equations (4.7) and (4.10), equation (4.10) must include at least on variable that is not in equation (4.7) (Maddala, 1983). I use three variables for this purpose. These exclusion restrictions variables are the components of vector $\underline{Z_{it}^{S0}}$, and they include height, weights, and a dummy for self-reported health status of the respondents. It is assumed that these variables influence the marriage search process. Another important identification condition of the system in equations (4.7) and (4.10) is that, Y_{it}^{0S}, the smoking response variable, of equation (4.7) is not allowed to feedback into the marriage specification of equation (4.10) (Maddala, 1983).

Unlike the OP model, the estimation of the OP-ED model involves evaluating multivariable outcomes with several integration points. In the literature, one of the methods doing this is the maximum simulated likelihood (MSL) (Train, 2003). Two popular methods for estimating the MSL are the Geweke-Hajivassiliou-Keane (GHK) simulator and the Gauss-Hermite quadrature/Adaptive quadrature simulator (GHQ). Both methods are based on MSL methods of Train (2009). Estimates of both methods are also identical. However, I use the GHK simulation method in this Chapter 5 because of its ease of implementation and speed (see Rodman; 2013).

The GHK simulator (Train, 2009; Roodman, 2011) uses multidimensional normal integrals and Halton draws. I devote a section of Chapter 5 to the derivation and description of MSL under

GHK simulator method. Under the GHK simulator, the likelihood function for the panel dataset of equations (4.7) and (4.10) is:

$$L = \prod_{i=1}^{N}\prod_{t=1}^{T}\left(pr(Y_{it}^{0S} = k, M_{it} = m)\right) \qquad m = \{0,1\}. \tag{4.11}$$

and the corresponding maximum simulated log-likelihood function is:

$$L(\theta) = \sum_{i=1}^{N}\log\left(\prod_{t=1}^{T} pr\left(Y_{it}^{0S} = k, M_{it} = m\right)\right) \qquad m = \{0,1\}. \tag{4.12}$$

where

$$pr\left(Y_{it}^{0S} = l, M_{it} = m\right) = \prod_{j=0}^{3} P\left(Y_{it}^{0S} = l, M_{it} = 0\right)^{(1-M_{it})Y_{it}^{OSj}} \times \prod_{j=0}^{3} P\left(Y_{it}^{0S} = l, M_{it} = 1\right)^{M_{it}Y_{it}^{OSj}} \tag{4.13}$$

with $j = \{0,1,2,3\}$, and $Y_{it}^{0Sj} = 1\ \{Y_{it}^{0S} = 1\}$ given the indicator function $1\{A\}$. According to Train (2009), to simulate and motivate the joint probability estimation of equations (4.7) and (4.10), one needs to impose an unobserved heterogenous structure on the error terms $\varepsilon_{it} = \left[\varepsilon_{it,Y}, \varepsilon_{it,M}\right]'$ on the system. This approach is also the one used by Bratti and Miranda (2010).[5] The full detail of this estimation procedure is also shown in Chapter 5. The variance covariance of ε_{it} structure is given by:

$$\Omega_{\varepsilon} = \begin{bmatrix} \omega_{YY} & \omega_{YM} \\ \omega_{MY} & \omega_{MM} \end{bmatrix}. \tag{4.14}$$

For identification of equation (4.13), I exclude the constant term of equation (4.10) as the base alternative from equation (4.10) (Train, 2009; Roodman, 2011). Following Roodman (2013), the constant term of equation (4.10) is this base alternative. Thus, the parameters estimated from

[5] The GHQ simulator (Bratti and Miranda, 2010) induces a relationship between the unobserved heterogeneous error terms $\varepsilon_{it,Y}$ and $\varepsilon_{it,M}$ by additional models: $\varepsilon_{it,Y} = \tau u_{it} + \aleph_{it}$, and: $\varepsilon_{it,M} = u_{it} + \omega_{it}$.

equation (4.12), the joint likelihood estimation likelihood of equations of (4.7) and (4.10), are all shown the vector parameter below:

$$\underline{\theta'} = \left(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \underline{\pi}, \underline{\delta}, \underline{\Omega_\varepsilon}, \underline{\mu_j}\right). \tag{4.15}$$

The derivation of drinking intensity models also follows the equations and estimation strategy that I described above in equations (4.1) to (4.15). That is, the estimates of the joint probabilities of drinking and marriage outcomes can be obtained by repeating the methodology used in equations (4.1) to (4.15) after substituting the alcohol consumption outcome variable for that of smoking. Hence, I substitute Y_{it}^{0d*} for Y_{it}^{0S*}, and Y_{it}^{0d} for Y_{it}^{0S} in equations (4.7) and (4.9) respectively. I repeat this exercise each time I identify a variable for smoking outcome.

4.3.3 The ZIOP model specification of smoking and drinking intensities

This section analyzes the Zero-inflated ordered probit (ZIOP) model. Like the OP and OP-ED models, the ZIOP model specification is based on the baseline regression. I conduct a test of goodness of fit between the ZIOP model and its correlated errors counterpart called the Zero-inflated ordered probit model correlated (ZIOPC) model in Chapter 5. I also match the ZIOP model to an ordered probit model with sample selection (OP-SS) model in Chapter 5.

The ZIOP model is a model of inflated or excess zeros, the third type of zeros identified in Chapter 1. The possibility of inflated zeros often arises if a survey question on smoking or drinking covers a short period. For example, it may be difficult to distinguish between respondents who have *never* smoked (coded with a "0" on the NLSY97) and infrequent smokers who have not smoked in the last 30 days (also coded with a "0" in the NLSY97) in a smoking survey question. Most surveys questions typically cover a short time periods for reasons that are beyond the scope of this study. For zero consumption, survey questions on smoking over a short period comprises respondents who have perfectly inelastic demand for tobacco ("abstention zeros"), smokers who

may consider switching to positive consumption if the conditions are right ("corner" solution), and infrequent smokers. Since the NLSY97 smoking questions cover 30 days, different types of zeros are all lumped together in each wave of the survey. The ZIOP model fits a model that distinguishes between these inflated zeros.

Put differently, the ZIOP model considers the possibility that zero consumptions can arise from two data generation processes (DGP). In the first case, a respondent must first decide whether to be a smoker or not. The zeros in the case belong to a DGP. Conditional on being a smoker (or being a participant), a respondent must decide the amount of tobacco product to consume, including zero consumption. A zero consumption that arises from this source belongs to a different DGP than the first DGP I just identified. Since the ZIOP model is a split between participation and consumption behavior, it is a double hurdle model of outright abstention (participation decision with zero outcome) and participation with zero consumptions (consumption decision with zero outcome).

Like in the previous sections, I analyze the ZIOP model of smoking intensities. I then generalize the model specifications to drinking intensities under the ZIOP model. The ZIOP model is a combination of a probit model and an ordered probit model. The propensity for participation is denoted with the latent variable S_{it}^{*} and it is described by:

$$S_{it}^{*} = \underline{\Pi}\underline{K_{it}} + u_{it,M}\,, \quad u_{it,M} \sim N(0,1) \tag{4.16}$$

where $S_{it} = \begin{cases} 1 & \text{if } S_{it}^{*} > 0 \\ 0 & \text{otherwise}\,. \end{cases}$

S_{it}^{*} denotes a latent variable outcome indicating the split between participants and nonparticipants in smoking activity. $u_{it,M}$ is the participation equation error term. $\underline{\Pi}$ is a *K X 1* parameter vector.

S is mapped to its latent variable equation S^* via the lower mathematical expression of equation (4.16). The expression in equation (4.16) is a probit model. $\underline{K_{it}}$ is a covariate of variables that determines an individual's decision on whether to participate or not. $\underline{K_{it}}$ is a covariate of socioeconomic and demographic variables that affects the probability of first marriage event. $\underline{K_{it}}$ includes most the socioeconomic and demographic variables whose summary statistics are shown in Table 5 of section 3.2. However, $\underline{K_{it}}$ does not include CPI of tobacco, income, CPI of alcoholic beverage (a complement), and the street price of marijuana (a substitute). These four variables have been previously identified in section 4.2.1 as economic factors that influence the demand for tobacco products. Non-inclusion of these variables in equation (4.16) makes sense because life-long nonparticipants have perfectly inelastic demand for tobacco product.

To model the consumption equation of the ZIOP model, I brought forward equation (4.7) of section 4.2.1:

$$Y_{it}^{0S*} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^*) + \theta_3 d_3(t - t_2^*) + +\underline{\pi}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}. \quad (4.17)$$

All the description and definitions of the variables in equation (4.17) are the same as they were previously described in section 4.2.1. The smoking ordinal response variable Y_{it}^{0S} describes the amount of tobacco product that an individual consumes conditional on participation. $\underline{X_{it}} = [K_{it}, R_{it}^{S0}]$. R_{it}^{S0} is a vector of the four economic variables that influence the amount of tobacco consumption conditional on participation. That is, R_{it}^{S0} vector is comprised of CPI of tobacco, income, CPI of alcoholic beverage (a complement), and the street price of marijuana (a substitute). Thus, these four variables serve to distinguish between abstention zeros and inflated zeros.

The derivation of the ZIOP model from the joint likelihood of equations (4.16) and (4.17) requires observing the smoking intensity variables, S_{it} and Y_{it}^{0S}. But S_{it} and Y_{it}^{0S} are not

individually observable. One needs to specify a third response variable through which S_{it} and Y_{it}^{0S} will be observable. Let $\tilde{y}_{it}^{S}$ be an ordinal intensity variable of smoking. S_{it} and Y_{it}^{0S} are observed through the criterion:

$$\tilde{y}_{it}^{S} = Y_{it}^{0S} * S_{it}. \tag{4.18}$$

Here, a positive value of $\tilde{y}_{it}^{S}$ is observed if, and only if $Y_{it}^{0S} > 0$ and $S_{it} = 1$. If $S_{it} = 0$, a zero value ('genuine' zero) of $\tilde{y}_{it}^{S}$ is observed. The probability of participation in equation (4.16) is given by[6]:

$$pr(S = 1|W) = pr(S^{*} > 0) = \Phi(W'\Pi). \tag{4.19A}$$

If the error terms in the participation equation (4.16) and consumption equation (4.17) are uncorrelated, the probability of observing the outcome variables in equation (4.17), including zero consumption conditional on participation is given by Maddala (1983):

$$\begin{aligned}
&\mathrm{P}(\tilde{y}^{S} = 0|\mathrm{W},\mathrm{V}) = \Phi[1 - \Phi(\mathrm{W}'\Pi)] + \Phi(V'\theta)\,\Phi(-V'\theta) \\
&\mathrm{P}(\tilde{y}^{S} = 1|\mathrm{Z},\mathrm{V}) = \Phi(V'\theta)[\Phi(\mu_1 - V'\theta) - \Phi(-V'\theta)] \\
&\mathrm{P}(\tilde{y}^{S} = 2|\mathrm{Z},\mathrm{V}) = \Phi(V'\theta)[\Phi(\mu_2 - V'\theta) - \Phi(\mu_1 - V'\theta) \\
&\mathrm{P}(\tilde{y}^{S} = 3|\mathrm{Z},\mathrm{V}) = \Phi(V'\theta)[1 - \Phi(\mu_2 - V'\theta)]
\end{aligned} \tag{4.19B}$$

where $V = (M_{it}, d_1, d_2, t, X_{it})$, $\underline{\theta'} = \left(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \underline{\pi}, \underline{\delta}, \underline{\Omega_{\varepsilon}}, \underline{\mu_j}\right)$, and $\Phi(.)$ is the cumulative distribution function (*cdf)* of a univariate normal distribution. Summing the individual likelihood functions across observations, the MLE of the ZIOP model is given by:

$$l(\underline{\hat{\theta}}) = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=0}^{J} h_{itj} \ln P(\tilde{y}^{S} = j|V, W; \underline{\theta}) \tag{4.20}$$

where $h_{ij} = \begin{cases} 1 \text{ if individual } i \text{ changes outcome} \\ 0 \text{ if otherwise } \end{cases}$ $i = 1,..,N.\ j = 0,1,2,3$

and $\underline{\hat{\theta}} = \max_{\underline{\theta}} l(\underline{\theta})$.

[6] From this point on, I dropped the subscripts to avoid repetitions

CHAPTER 5: ECONOMETRICS SPECIFICATION

I showed the derivations of the likelihood functions of the ordered probit (OP), ordered probit with endogenous dummy (OP-ED), and the zero-inflated ordered probit model (ZIOP) models in Chapter 4[7]. I used the equations and likelihood functions of these models without repeating their derivations in this chapter. I show the implementation of the maximum simulated likelihood (MSL) in this chapter. As in Chapter 4, I only show the derivation of the smoking intensity models in this chapter. I then generalize the derivations to the drinking intensity models. In addition to the implementation of the MSL method, I show the derivations of the marginal effects of zero-inflated ordered probit (ZIOP) and zero-inflated ordered probit correlated (ZIOPC) models.

5.1 The maximum simulated likelihood (MSL) method of estimating the OP-ED model

This section develops the derivation and implementation of the maximum simulated likelihood (MSL) method, a method that is often employed in the literature to estimate the OP-ED model. Without loss of generality, the relevant derivations of smoking intensities from Chapter 4 are:

$$Y_{it}^{0S} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^*) + \theta_3 d_3(t - t_2^*) + + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0s}}. \quad (5.1)$$

$$M_{it}^* = \delta W_{it} + \varepsilon_{it,M} \quad (5.2)$$

[7] See equations (4.9), (4.12) and (4.20) of Chapter 4 for the likelihood function of the OP, OP-ED and ZIOP models respectively.

where $M_{it} = 1$ if $M_{it}^{*} > 0$ or $M_{it} = 0$ otherwise. Equation (5.1) is the baseline OP model, and it is the same as equation (4.7) of Chapter 4. Equation (5.2) is the marriage selection equation, and it is the same as equation (4.10) of Chapter 4.

Equations (5.3) – (5.7) below are the same as equations (11) - (15) in Chapter 4. Under the GHK simulator (Train, 2009; Roodman, 2011) assumptions in Chapter 4, I derive that the likelihood functions in equations (5.3) – (5.5) below:

$$L = \prod_{i=1}^{N}\prod_{t=1}^{T}\left(pr(Y_{it}^{0S} = k, M_{it} = m)\right) \qquad m = \{0,1\}. \tag{5.3}$$

The corresponding maximum simulated log-likelihood function is:

$$L(\theta) = \sum_{i=1}^{N}\log\left(\prod_{t=1}^{T} pr\left(Y_{it}^{0S} = k, M_{it} = m\right)\right) \qquad m = \{0,1\} \tag{5.4}$$

$$pr\left(Y_{it}^{0S} = l, M_{it} = m\right) = \prod_{j=0}^{3} P\left(Y_{it}^{0S} = l, M_{it} = 0\right)^{(1-M_{it})Y_{it}^{OSj}} \times \prod_{j=0}^{3} P\left(Y_{it}^{0S} = l, M_{it} = 1\right)^{M_{it}Y_{it}^{OSj}} \tag{5.5}$$

where $j = \{0,1,2,3\}$, and $Y_{it}^{0Sj} = 1\,\{Y_{it}^{0S} = 1\}$ given the indicator function $1\{A\}$. The variance covariance of $\varepsilon_{it} = \left[\varepsilon_{it,Y^{0S}}, \varepsilon_{it,M}\right]'$ structure is given by:

$$\Omega_{\varepsilon} = \begin{bmatrix} \omega_{YY} & \omega_{YM} \\ \omega_{MY} & \omega_{MM} \end{bmatrix}. \tag{5.6}$$

The parameters estimated in equation (5.4) are shown in the vector parameter below:

$$\underline{\theta'} = \left(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \underline{\pi}, \underline{\delta}, \underline{\Omega_{\varepsilon}}, \underline{\mu_j}\right). \tag{5.7}$$

Without loss of generality, one can show different components of the joint distributions of equation (5.5). For example, if an individual is married, and has not smoked in the last 4 weeks, the relevant equation one needs to evaluate in equation (5.5) is:

$$P\left(Y_{it}^{0S} = 0, M_{it} = 1\right) = P\left(Y_{it}^{0S*} < 0, M_{it}^{*} > 0\right)$$

$$= P\left(-\infty - H_{it}^{Y} < \varepsilon_{it,Y} < -H_{it}^{Y}, \varepsilon_{it,M} > H_{it}^{M}\right) \tag{5.8}$$

$$= P(-\infty - H_{it}^{Y} < \varepsilon_{it,Y} < -H_{it}^{Y} \mid \varepsilon_{it,M} > H_{it}^{M}) \times P(\varepsilon_{it,M} > H_{it}^{M})$$

where $H_{it}^{Y} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^*) + \theta_3 d_3(t - t_2^*) + + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0S}}$ and $H_{it}^{M} = \delta W_{it}$.

Similar probability expressions can be shown for other joint distributions in equation (5.5).

Following Bratti and Miranda (2010), Roodman (2010), and Train (2009), I induce a relationship between the unobserved heterogeneous error terms $\varepsilon_{it,Y}$ and $\varepsilon_{it,M}$ by an additional model:

$$\varepsilon_{it,M} = u_{it} + \omega_{it} \tag{5.9A}$$

$$\varepsilon_{it,Y} = \tau u_{it} + \aleph_{it}. \tag{5.9B}$$

τ is a free parameter or a factor loading parameter, and $\aleph_{it}, \omega_{it}, u_{it} \sim N(0,1)$. The variance-covariance matrix of $\aleph_{it}$ and ω_{it} is:

$$\begin{bmatrix} \aleph_{it} \\ \omega_{it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right). \tag{5.10}$$

Let an indicator $I_{ij} = 1$ if $Y_{it}^{0S} = j$, and $I_{ij} = 0$ for $j = 0,1,2,3$. The simulated likelihood function (Bratti and Miranda, 2010) of equation (5.3) in terms of equations (5.9A) and (5.9B) can be explicitly written as:

$$L = \int_{-\infty}^{\infty} \sum_{t=1}^{T} \sum_{j=1}^{J} I_{ij} \Phi_j \{M_{it}\, \Phi(\delta W_{it} + u_{it}) + (1 - M_{it})[1 - \Phi(\, \delta W_{it} + u_{it})]\} \phi(u_{it}) du_{it} \tag{5.10}$$

with

$$\Phi_j = \begin{cases} 1 - \Phi(V'\Gamma - \mu_1 + \tau u_{it}) & \text{if } j = 1 \\ \Phi\big(V'\Gamma - \mu_{j-1} + \tau u_{it}\big) - \Phi\big(V'\Gamma - \mu_J + \tau u_{it}\big) & \text{if } 1 < j < J - 1 \\ \Phi\big(V'\Gamma - \mu_{J-1} + \tau u_{it}\big) & \text{if } j = J \end{cases} \tag{5.11}$$

where $\underline{\Gamma} = \big(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \underline{\pi}\big)$ and $V = (M_{it}, d_1, d_2, t, X_{it})$. The vector of θ of equation (5.4) can thus be expressed as $\underline{\theta'} = \Big(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \underline{\pi}, \underline{\delta}, \underline{\Omega_\varepsilon}, \underline{\mu_j}, \rho\Big)$.

The correlation coefficient between the error terms of the baseline regression in equation (5.1) and the marriage self-selection equation in equation (5.2) is denoted by ρ. This correlation coefficient can be obtained by:

$$\rho = \frac{cov\left(\varepsilon_{it,Y},\ \varepsilon_{it,M}\right)}{\sqrt{var\ (\varepsilon_{it,Y}\ )}\sqrt{var(\varepsilon_{it,M})}} \tag{5.12}$$

where

$$cov\left(\varepsilon_{it,Y},\ \varepsilon_{it,M}\right) = E\left[\left(\varepsilon_{it,Y} - E\left(\varepsilon_{it,Y}\right)\right)\left(\varepsilon_{it,M} - E\left(\varepsilon_{it,M}\right)\right)\right] = \tau \tag{5.13A}$$

$$var\left(\varepsilon_{it,Y}\right) = var(\tau u_{it}) + var(\aleph_{it}) + 2cov(\tau u_{it}, \aleph_{it}) = \tau^2 + 1 \tag{5.13B}$$

$$var\left(\varepsilon_{it,M}\right) = var(u_{it}) + var(\omega_{it}) + 2cov(u_{it} + \omega_{it}) = 2. \tag{5.13C}$$

Plugging equation (5.12) into equation (5.13):

$$\rho = \frac{\tau}{\sqrt{2(\tau^2 + 1)}}. \tag{5.14}$$

I use the t-test test of this correlation coefficient as one the tests of fit between the OP and OP-ED models in Chapter 6. If $\rho = 0$, then $\varepsilon_{it,Y}$ and $\varepsilon_{it,M}$ are independent. If this test statistic is significant under the null hypothesis of $\rho = 0$, then I infer that marriage is endogenous in the model and that it has a causal effect on smoking. In other words, the OP-ED model is the correct specification and it is a superior fit than the OP model.

The final step of the implementation of the OP-ED model involves reparameterization of the expressions in (5.3) – (5.8) and equation (5.10) for all the possible joint distributions of the baseline regression and the marriage self-selection equation (see Train, 2009; Bratti and Miranda, 2010). For example, the joint distribution of an unmarried smoker who smokes less than once in a week, $P(Y_{it} = 1, M_{it} = 0)$, is:

$$P(Y_{it} = 1, M_{it} = 0) = P\left(\aleph_{it} < \left(\frac{H_{it}^{Y} + \tau u_{it}}{\sqrt{(\tau^2 + 1)}}\right) \Big| \omega_{it} < \frac{-H_{it}^{M} + u_{it}}{\sqrt{2}}\right) \times P\left(\omega_{it} < \frac{-H_{it}^{M} + u_{it}}{\sqrt{2}}\right)$$

$$= \Phi\left(\frac{-H_{it}^{M}+u_{it}}{\sqrt{2}}\right) \times \int_{-\infty}^{\frac{-H_{it}^{M}+u_{it}}{\sqrt{2}}} \Phi\left(\frac{H_{it}^{Y}+u_{it}}{\sqrt{(\tau^2+1)}}\right)\phi(u_{it})du_{it} \quad (5.15)$$

where $H_{it}^{Y} = \beta_1 M_{it} + \beta_2 M_{it} * t + \pi_1 * t + \gamma' X_{it}$, and $H_{it}^{M} = \delta W_{it}$.

Using the probabilities of each joint distribution above, one can calculate the simulated GHK probability in the following steps as shown in Train (2009):

1. Calculate $P\left(\omega_{it} < \frac{-H_{it}^{M}+u_{it}}{\sqrt{2}} \mid \tilde{\alpha}_i^r\right) = \Phi\left(\frac{-H_{it}^{M}+u_{it}}{\sqrt{2}}\right)$.

2. Draw a value of ω_{it}^{M}, labeled $\omega_{it}^{M,z}$, from a truncated standard normal truncated at $\frac{-H_{it}^{M}+u_{it}}{\sqrt{2}}$ and this draw can be obtained in the following way:

a. Draw a μ_1^z , as the $z^{t\square}$ element of the first Halton Sequence of length M.

b. Calculate $\omega_{it,}^{z} = \Phi^{-1}\left((1-\mu_1^r)\Phi(-\infty) + \mu_1^r \Phi\left(\frac{-H_{it}^{M}+u_{it}}{\sqrt{2}}\right)\right) = \Phi^{-1}\left(\mu_1^r \Phi\left(\frac{-H_{it}^{M}+u_{it}}{\sqrt{2}}\right)\right)$.

3. Calculate $P\left(\aleph_{it} < -\left(\frac{H_{it}^{Y}+\tau u_{it}}{\sqrt{(\tau^2+1)}}\right) \mid \omega_{it} = u_{it}^{M,m}, \tilde{\alpha}_i^r\right) = \Phi\left(\frac{-H_{it}^{Y}-\tau u_{it}}{\sqrt{(\tau^2+1)}}\right)$.

4. Draw a value of $\aleph_{it}^{Y}$, labeled $\aleph_{it}^{Y,z}$ from a truncated standard normal truncated at $-\left(\frac{H_{it}^{Y}+u_{it}}{\sqrt{(\tau^2+1)}}\right)$. This draw is obtained as follows:

a. Draw a μ_2^z, as the $z^{t\square}$ element of the second Halton Sequence of length Q.

b. Calculate $\aleph_{it}^{z} = \Phi^{-1}\left((1-\mu_2^r)\Phi(-\infty) + \mu_2^r \Phi\left(\frac{-H_{it}^{Y}-\tau u_{it}}{\sqrt{(\tau^2+1)}}\right)\right) = \Phi^{-1}\left(\mu_2^r \Phi\left(\frac{-H_{it}^{Y}-\tau u_{it}}{\sqrt{(\tau^2+1)}}\right)\right)$.

5. The simulated probability for $z^{t\square}$ draw of ε_{it}^{M} and ε_{it}^{Y} is computed as: $P(Y_{it}=1, M_{it}=0) = \Phi\left(\frac{-H_{it}^{M}+u_{it}}{\sqrt{(\tau^2+1)}}\right) \times \Phi\left(\frac{-H_{it}^{Y}-\tau u_{it}}{\sqrt{(\tau^2+1)}}\right)$.

6. Repeat steps 1 to 5 for $z = 1,2,..,z$.

7. Thus, the simulated probability is $\tilde{P}(Y_{it}=1, M_{it}=0) = \frac{1}{Z}\sum_{z=1}^{Z} P(Y_{it}=1, M_{it}=0)^z$.

5.2. The marginal effects of the ZIOP and ZIOPC models

I show the marginal effects of the ZIOP and ZIOPC models in this section. The ZIOPC model is a ZIOP model with the possibility that the error terms are correlated. I provide a full derivation of the ZIOPC model from the ZIOP model later in this section.

The marginal effect of an outcome variable in an ordered probit model considers the probability that the event described by the outcome variable will occur. The unconditional mean of an ordered probit model, $[Y|X]$, does not exist (Greene and Hensher, 2010; Harris and Zhao, 2007). Since all the models in this study are based on variations of the ordered probit model, their estimates are not meaningful. As a result, one needs to rely on marginal effects for meaningful economic analyses and interpretations of the OP, the OP-ED, the SSOP and the ZIOP models.

From equation (4.16) in Chapter 4, the propensity for participation is denoted by the latent variable S_{it}^{*} and it is described by:

$$S_{it}^{*} = \underline{\Pi}\underline{K_{it}} + u_{it,M}\,, \quad u_{it,M} \sim N(0,1) \tag{5.16}$$

where $S_{it} = 1$ if $S_{it}^{*} > 0$ and $S_{it} = 0$ otherwise. All the description and definitions of the variables in equation (5.16) are as previously described in section 4.2.1. To model the consumption equation of the ZIOP model, I bring forward equation (4.7) of section 4.3.1:

$$Y_{it}^{0S*} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^{*}) + \theta_3 d_3(t - t_2^{*}) + +\underline{\pi}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}} \tag{5.17}$$

All the description and definitions of the variables in equation (5.17) are also as previously described in section 4.3.1.

The standard assumptions about the thresholds constants of the ordered probit model are:

$$Y_{it}^{0S} = \begin{cases} 0 & \text{if } Y_{it}^{0S*} \le 0 \\ 1 & \text{if } 0 < Y_{it}^{0S*} \le \mu_1 \\ 2 & \text{if } \mu_1 < Y_{it}^{0S*} \le \mu_2 \\ 3 & \text{if } \mu_2 < Y_{it}^{0S*} \end{cases}. \tag{5.18}$$

The derivation of the ZIOP model from the joint likelihood of equations (5.16) and (5.17) requires observing the smoking intensity variable, S_{it} and Y_{it}^{0S}. But S_{it} and Y_{it}^{0S} are not individually observable. Let the $\tilde{y}_{it}^{S}$ be an ordinal intensity variable of smoking. S_{it} and Y_{it}^{0S} are observed through the criterion:

$$\tilde{y}_{it}^{S} = Y_{it}^{0S} * S_{it}. \tag{5.19}$$

Here, a positive value of $\tilde{y}_{it}^{S}$ is observed if and only if $Y_{it}^{0S} > 0$ and $S_{it} = 1$. If $S_{it} = 0$, a zero value of $\tilde{y}_{it}^{S}$ is observed. The probability of participation in equation (5.16) is:

$$pr(S = 1|K) = pr(S^{*} > 0) = \Phi(K'\Pi). \tag{5.20}$$

If the error terms in the participation equation (5.16) and consumption equation (5.17) are uncorrelated, then the probabilities of observing the outcome variables in equation (5.17), including zero consumption, conditional on participation are:

$$\mathrm{P}(\tilde{y}^{S} = 0|\mathrm{K},\mathrm{V}) = \Phi[1 - \Phi(\mathrm{K}'\Pi)] + \Phi(V'\theta)\Phi(-V'\theta) \tag{5.21A}$$

$$\mathrm{P}(\tilde{y}^{S} = 1|\mathrm{K},\mathrm{V}) = \Phi(K'\theta)[\Phi(\mu_1 - V'\theta) - \Phi(-V'\theta)] \tag{5.21B}$$

$$\mathrm{P}(\tilde{y}^{S} = 2|\mathrm{K},\mathrm{V}) = \Phi(K'\theta)[\Phi(\mu_2 - V'\theta) - \Phi(\mu_1 - V'\theta) \tag{5.21C}$$

$$\mathrm{P}(\tilde{y}^{S} = 3|\mathrm{K},\mathrm{V}) = \Phi(K'\theta)[1 - \Phi(\mu_2 - V'\theta)] \tag{5.21D}$$

where $V = (M_{it}, d_1, d_2, t, X_{it})$, $\underline{\theta'} = \left(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \underline{\pi}, \underline{\delta}, \underline{\Omega_{\varepsilon}}, \underline{\mu_j}\right)$, and $\Phi(.)$ is the cumulative distribution function (*cdf)* of a univariate normal distribution. Summing the individual likelihood functions across observations, the MLE of the ZIOP model is given by:

$$l(\underline{\hat{\theta}}) = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=0}^{J} h_{itj} \ln P(\tilde{y}^{S} = j|V, K; \underline{\theta}) \tag{5.22}$$

where $h_{ij} = \begin{cases} 1 \text{ if individual } i \text{ changes the outcome} \\ 0 \quad \text{if otherwise } \ i = 1,..,N.\ j = 0,1,2,3 \end{cases}$

and $\underline{\hat{\theta}} = \max_{\underline{\theta}} l(\underline{\theta})$.

The expression for the marginal effects of the ZIOP model in equations (5.21A-D) is given by:

$$\frac{\partial P(y=0|V)}{\partial V} = [\Phi(-\theta'V) - 1]\phi(\Pi'W)\Pi - \Phi(\Pi'W)(\theta'V)\theta \quad (5.22A)$$

$$\frac{\partial P(y=j|V)}{\partial V} = \left[\Phi\left(\mu_j - \theta'V\right) - \left(\mu_{j-1} - \theta'V\right)\right]\phi(\Pi'W) \quad (5.22B)$$

$$+ \left[\Phi\left(\mu_{j-1} - \theta'V\right)\phi(\Pi'W) - \Phi\left(\mu_j - \theta'V\right)\phi(\Pi'W)\right]$$

To motivate the derivation of the marginal effects of the ZIOPC model, I consider the distributional assumption of the error terms covariance matrices from the error terms of the participation and consumption equations in equations (5.16) and (5.17). Let ρ denote the correlation coefficient between the error terms of these equations. The bivariate normal distribution assumption of the covariance matrix is:

$$\begin{pmatrix} \varepsilon_{it,Y^{0s}} \\ u_{it,M} \end{pmatrix} \sim BIVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]. \quad (5.23)$$

Harris and Zhao (2007), and Greene and Hensher (2010) show that the joint probability distribution of the ZIOP model with correlated error terms is given by:

$$P\,(\tilde{y}^S = 0|\mathrm{K}, V) = [\,(-\mathrm{K}\Pi)] + \Phi_2(\mathrm{K}\Pi, V\theta; -\rho) \quad (5.24A)$$

$$P(\tilde{y}^S = j|K,V) = \Phi_2\left(\Pi'K, \mu_j - V\varphi; -\rho\right) - \Phi_2\left(\Pi'K, \mu_{j-1} - V\varphi; -\rho\right) \quad (5.24B)$$

$$P(\tilde{y}^S = J|K,V) = \Phi_2\left(\Pi'K, V\varphi - \mu_{J-1}; -\rho\right) \quad (5.24C)$$

$j = 1, \dots, J-1,\ V = (M_{it}, d_1, d_2, t, X_{it}),\ \underline{\theta'} = \left(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \underline{\pi}, \underline{\delta}, \underline{\Omega_\varepsilon}, \underline{\mu_j}, \rho\right),$ and $\Phi(.)$ is the cumulative distribution function (*cdf)* of a univariate normal distribution. $\Phi_2(.)$ is the cumulative distribution function of a bivariate normal distribution. ρ is the correlation between $\varepsilon_{it,Y^{0s}}$ and $u_{it,M}$.

I use the t-test test of the correlation coefficient of the error terms of the ZIOPC model as one of the tests of goodness of fit between the ZIOP and ZIOPC models. If $\rho = 0$, then $u_{it,Y}$ and $u_{it,M}$ are independent. Under the null hypothesis $\rho = 0$, if the test statistic is significant, then I infer that the ZIOPC model is the appropriate model. Otherwise the ZIOP model is the appropriate model.

The expression for the marginal effects of equations (5.23) is given by the expression (see Greene and Hensher, 2010):

$$\frac{\partial P(y=0|V)}{\partial V}$$

$$= \left[\Phi\left(\frac{-\theta'V + \rho\Pi'K}{\sqrt{(1-\rho^2)}}\right) - 1\right]\phi(\Pi'W)\Pi - \Phi\left(\frac{\Pi'K - \rho\theta'V}{\sqrt{(1-\rho^2)}}\right)\phi(\theta'V)\frac{\partial P(y=j|V)}{\partial V}$$

$$= \left[\Phi\left(\frac{\mu_j - \theta'V + \rho\Pi'W}{\sqrt{(1-\rho^2)}}\right) - \Phi\left(\frac{\mu_{j-1} - \theta'V + \rho\Pi'W}{\sqrt{(1-\rho^2)}}\right)\right]\phi(\Pi'W)\Pi$$

$$+ \begin{bmatrix} \phi(\mu_{j-1} - \theta'V)\Phi\left(\frac{\Pi'W + \rho(\mu_{j-1} - \theta'V)}{\sqrt{(1-\rho^2)}}\right) \\ -\phi(\mu_j - \theta'V)\Phi\left(\frac{\Pi'W + \rho(\mu_j - \theta'V)}{\sqrt{(1-\rho^2)}}\right) \end{bmatrix}\theta. \qquad (5.25)$$

5.3 The Ordered probit model with sample selection (SSOP) model, and sample attrition

In this section, I present the log-likelihood of the Heckman ordered probit model with sample selection (SSOP). I use the SSOP model to address sample attrition and survivorship bias in this study. A comprehensive derivation and analysis of the SSOP model is in De Luca and Perotti (2011).

Sample selection can pose a serious estimation issue if attrition is not random. In the NLSY97 sample, I removed 1,561 (or 17.37% of the original sample) respondents from the smoking sample in Table 3.3 of Chapter 3 due to their 'non-interview' status. Similarly, I also

removed 1,843 (or 20.51%) respondents from the drinking sample in Table 3.4 of Chapter 3 due to their 'non-interview' status. Reasons for 'non-interview' status include death, incarceration, refusals for interviews, and non-locatability.

If this attrition process is not random, the estimates of the OP, the OP-ED, ZIOP, and the ZIOPC models will suffer from upward survivorship bias because the 'well-behaved' respondents in the sample are more likely to be married, and thus smoke and drink less. To address this survivorship bias issue in the smoking equation, one needs variables that are correlated with the probability of the selection equation but uncorrelated with the probability of smoking. Similarly, the drinking equation needs to include variables that are uncorrelated with the probability of drinking in the selection equation.

These unique variables are included with other socio-economic characteristics of the respondents to predict the probability of successful conduct of an interview in the selection equation. Following De Luca and Perotti (2011), I use experience, age, gender, and educational attainment of the interviewers for these unique variables in the selection equation.

An ordered probit model with sample selection (SSOP) model can be modeled with the equations below:

$$D_{it}^{*} = \underline{\omega}\underline{L_{it}} + u_{it,L}, \quad u_{it,L} \sim N(0,1) \tag{5.26}$$

$$Y_{it}^{0S} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^{*}) + \theta_3 d_3(t - t_2^{*}) + +\underline{\pi}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}} \tag{5.27}$$

Equation (5.26) is the selection equation. $D_{it} = 1$ if Y_{it}^{0S} is observed, otherwise $D_{it} = 0$. Equation 5.27 is the baseline regression equation. The joint probability distributions of these equations are:

$$Pr(D = 0|\mathrm{L}) = 1 - \Phi(\underline{\omega}\underline{L_{it}}) \tag{5.28A}$$

$$P(\tilde{y}^{S} = j|K,V) = \Phi_2\big(\underline{\omega}L, \mu_j - V\varphi; -\rho\big) - \Phi_2\big(\underline{\omega}L, \mu_{j-1} - V\varphi; -\rho\big) \tag{5.28B}$$

$$P(\tilde{y}^{S} = J|K,V) = \Phi_2\big(\underline{\omega}L, V\varphi - \mu_{J-1}; -\rho\big) \tag{5.28C}$$

where $j = 1, \ldots, J$. Equation (5.28) is an Heckman ordered probit model with sample selection (SSOP) (De Luca and Perotti (2011). It should be noted that $(\varepsilon_{it,Y^{os}}, u_{it,L})$ is distributed as bivariate normal distribution. If statistical test of the null hypothesis $\rho = 0$ is rejected, then I infer that the OP model is the superior of the SSOP and OP models.

CHAPTER 6: MODEL SELECTION AND HYPOTHESIS TESTING

The maximum likelihood and the marginal effects of the OP, the OP-ED, the ZIOP, the ZIOPC, and the SSOP models have been derived and examined in Chapters 4 and 5. To establish the appropriate model, I use the t-test to test whether the covariance of the error terms of the OP and OP-ED, OP and SSOP, and ZIOP and ZIOPC are equal to zero. I also test whether the correlation coefficients (ρ) between these pairs are equal to zero. Just as correlation does not imply causality, the tests between covariances of the error terms are more powerful than those between the correlation of the error terms.

In addition to the tests above, I also employ information-based criteria (AIC and BIC), and the Vuong (1989) test to determine the appropriate model between the model pairs in Figure 6.1.

6.1 Model selection pairs

Generally, the OP and OP-ED models can fit abstention zeros. As a reminder, the OP-ED model considers the possibility of self-selection in the marriage search process. ZIOP and ZIOPC models can fit inflated or excess zeros, while the SSOP model can fit survivorship bias. In Figure 6.1 below, I pair the models by the types of zero that I identified in Chapter 1, section 4.2 of Chapter 4, and section 5.3 of Chapter 5. I test these paired models in four steps. First, in the horserace 1, I test the OP and the OP-ED models against each other. In this horserace, I test whether the marriage selection process is random. Here, I tag the appropriate model of this horserace 1 as 'winner' 1 (W1). Second, in the horserace 2, I test the appropriate model between the winner 1 and SSOP model pair. In this case, I test if there is a survivorship bias in the underlying dataset. I name the winner of this horserace as 'Winner" 2 (W2). Third, in horserace 3, I compare the two

zero-inflated models, the ZIOP and ZIOPC pair. In this horserace, I test whether the error terms are correlated in the zero-inflated models. In this case, I tag the winner of this horserace as 'winner' 3 (W3). In the final horserace, I test for the appropriate model between W2 and W3.

In Figure 6.1, I show the model pairs, the steps, and the appropriate model questions. I present the results of these horseraces in section 6.4.

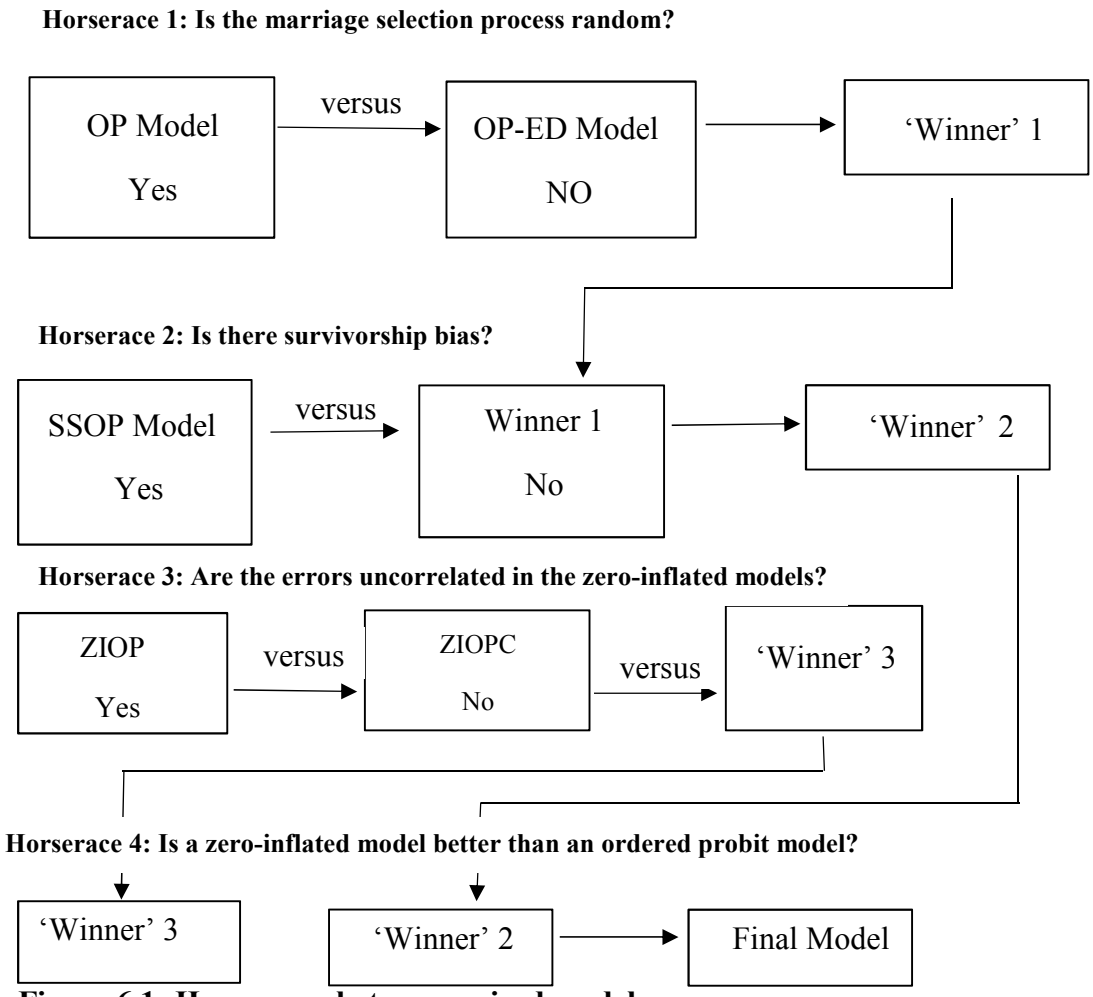

Figure 6.1: Horseraces between paired models

6.2 Model selection criteria tests

In this section, I present brief discussions and derivations of the goodness of fit tests and model selection criteria. Here, I examine the asymptotic standard normal tests of the covariances

of the error terms of model pairs (where applicable). I also analyze information-based criteria (AIC and BIC), and the Vuong (1989) test in this section.

6.2.1 The asymptotic standard normal tests of the covariances and correlation coefficients between the error terms

I use the t-test to test whether the covariances of the error terms between the OP and OP-ED models, the SSOP and OP models, and the ZIOP and ZIOPC models are equal to zero. I also use a second t-test to test whether the correlation (ρ) between the standard errors of the OP and OP-ED models, the SSOP and OP models, and the ZIOP and ZIOPC models are equal to zero. While the estimates of the covariances and correlations between the error terms are both from the maximum likelihood estimations, the former test is more powerful than the latter.

6.2.2 The information-based model selection criteria

Information-based model selection criteria are used to choose between nested and non-nested paired models. In this case, I use the Akaike (1973) information criterion (AIC) and the Bayesian information criterion (BIC). The AIC and BIC statistic are respectively given by:

$$AIC = -2l(\theta) + k \tag{6.1}$$

$$BIC = -2l(\theta) + (\ln N)k \tag{6.2}$$

where k is the number of parameters, N is the sample size, and $l(\theta)$ is the log-likelihood estimate. The AIC and BIC do not compare estimates of a model to some baseline null hypothesis. Under information-based model selection criteria, the model with the smaller information criterion is considered the better fit.

6.2.3 The Vuong test (Vuong, 1989)

The Vuong test (Vuong, 1989) has theoretical justifications for choosing between non-nested models (see Harris and Zhao, 2007; Humphreys, 2013). This test is based on a transformed log-likelihood values. It has been used in the context of model selection between the ZIOP and OP

models (Greene 2003; Harris and Zhao, 2007). The Vuong test is specifically designed for non-nested models (see Humphreys, 2013).

The construction of the Vuong test (1989) statistic and its associated critical value is as follows. Using the notations in equations (4.16) to (4.20), let $f_1(\tilde{y}_{it}^s|V_{it}, W_{it})$ be the predicted probability using the OP model. So, the OP is the model in the numerator of the expression in equation (6.3) below. Similarly, let $f_2(\tilde{y}_{it}^s|V_{it}, W_{it})$ be the predicted probability of the ZIOP model. So, the ZIOP is the model in the denominator of the expression in equation (6.3). The Vuong statistic is thus:

$$e_i = \log\left(\frac{f_1(\tilde{y}_{it}^s|V_{it}, W_{it})}{f_2(\tilde{y}_{it}^s|V_{it}, W_{it})}\right). \tag{6.3}$$

The transformed Vuong statistics, v, is given by:

$$v = \frac{\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} e_i\right)}{\sqrt{1/N \sum_{i=1}^{N}(e_i - \bar{e_i})^2}}. \tag{6.4}$$

The null hypothesis of the Vuong test (1989) is $E(e_i) = 0$. v in equation (6.4) has a standard normal distribution (see Vuong, 1989). The critical value of the Vuong test statistic at the 5% level of significance is $v = \pm 1.96$. Since this test is bidirectional, $|v| < 1.96$ favors neither model. OP fits the data better if $v > 1.96$, but the ZIOP is the superior model if $v < -1.96$. For the Vuong test, the null hypothesis is that the OP and the ZIOP models are equally close to the true data generating process, against the null hypothesis that one of them is more appropriate for the data generating process.

6.3 Goodness of fit results

In this section, I conduct goodness of fit tests in the order shown in Figure 6.1. In Tables 6.1 and 6.2 below, I present the results of these paired models. While the OP model conditions on

demographic variables and the four economic variables, the OP-ED model conditions on all the variables in the OP model plus six personal trait variables (height, weight, and four dummy variables for self-reported health status) of the respondents. But the ZIOP and ZIOPC models condition on participation and consumptions variables. The goodness of fit tests are shown in Tables 6.1 through 6.4. For smoking intensities, Tables 6.1 and 6.2 presents the goodness of fit tests results between the OP, the OP-ED, the SSOP, and the ZIOPC, and the ZIOP models. Table 6.3-6.4 show similar goodness of fit results for drinking intensities.

In the Tables below, "n/a" implies that the test is not applicable or appropriate for that model pair. All the tests are conducted at the 5% significance levels.

6.3.1 Smoking model selection criteria and tests of goodness of fit

I start with horserace 1 using the smoking sample. The first test of fit in Table 6.1 is between the OP and OP-ED model pair. For example, in columns 1 and 2, the estimate of the covariance of the error terms between these models, $\text{cov}(\varepsilon_{it,Y^{OP}}, \varepsilon_{it,Y^{OP-ED}})$, is 0.124, while the p-value under the null hypothesis that value of this covariance is zero is 0.326. Also, the estimate of ρ is 0.056, and the p-value of the null hypothesis that $\rho = 0$ is 0.134. I fail to reject either of the null hypotheses of these two tests. Thus, these two tests imply that marriage is not endogenous in the model, and that OP-ED model does not provide a superior fit than the OP model. As the most parsimonious model, the OP is chosen as the best fit based on these criteria. for the OP is a better specification than the OP-ED based on this criterion.

Table 6.1 Smoking Model: Selection Criteria and Test of Goodness of Fit for Horseraces 1 & 2

	Horserace 1		Horserace 2	
	OP Model (1)	OP-ED Model (2)	OP Model (3)	SSOP Model (4)
Covariance (p-value)	-	0.124 (0.326)	-	1.874 (0.175)
ρ (p-value)	-	0.056 (0.134)	-	0.025 (0.864)
AIC	43,406.440	50,801.870	43,406.440	35,049.170
BIC	43,668.980	51,318.250	43,668.980	35,540.840

Turning my attention to the other tests to supplement the correlation and covariance between the error terms tests, I find that values of the AIC and BIC information-based criteria of the OP model is less those of the OP-ED model, and this implies that the OP model is more appropriate than the OP-ED under the AIC and BIC information criteria.

Finally, since the OP and OP-ED models are overlapping models, the Vuong test is not applicable in this case (see Silva et al., 2014; Wilson 2015). In conclusion, the goodness of fit tests between the OP and the OP-ED model show that the OP model is the superior model as the smoking sample does not suffer from marriage self-selection bias. That is, the marriage selection process is random.

Next, I conduct horserace 2, the test for sample attrition (see Figure 6.1). In this case, sample attrition may cause survivorship bias if the attrition is not random. Just as in the test of specification between the OP and OP-ED models, the correlation coefficient and covariance of the error terms between the OP and Heckman sample selection ordered probit (SSOP) model pair are the most powerful tests. These tests are shown in columns 3 and 4 of Table 6.1 above.

The null hypothesis in this case is that the OP model does not suffer from survivorship bias problems. Specifically, under this null hypothesis, the p-values of covariance and correlation coefficient of the OP and SSOP error terms are respectively 0.175 and 0.864. In this case, I fail to reject the null hypothesis that the OP model does not suffer from sample attrition problems. But the AIC and BIC do not support this conclusion as both are lower to that of the SSOP model. But I consider the OP was the 'winner' of the horserace 2 because ρ and the covariance of the error terms are formal statistical tests whereas the AIC and BIC provides rule of the thumb.

Next, I conduct horseraces 3 and 4 for the smoking sample in Table 6.2. I conduct horserace 3 to determine the superior model between the two zero-inflated models – the ZIOP and ZIOPC. These results are shown in columns 1 and 2 of Table 6.2 below.

Table 6.2. Smoking Model: Selection Criteria and Test of Goodness of Fit for Horseraces 3 & 4

	Horserace 3		Horserace 4	
	ZIOPC Model (1)	ZIOP Model (2)	ZIOP Model (3)	OP Model (4)
Covariance (p-value)	-0.6341(0.452)	-	n/a	n/a
ρ (p-value)	-0.002 (0.192)	-	n/a	n/a
AIC	36,022.82	36,024.30	36,024.30	43,406.44
BIC	35,925.25	35,916.51	35,916.51	43,668.98
Vuong Test (p-value)	n/a	n/a	-22.21 (0.000)	n/a

Under the null hypothesis that the ZIOP model is the better model, the p-values of the covariance and correlation coefficients of the error terms of the ZIOPC and ZIOP models are respectively 0.452 and 0.192. These tests imply that I fail to reject the null hypotheses that neither covariance and ρ between the two error terms is equal to zero, and I infer that the ZIOP is the appropriate zero-inflated model. However, except for the AIC criterion, I also infer that, all the other specification tests support the superiority of the ZIOP over the ZIOPC model in Table 6.2.

In the final model selection, horserace 4 test between the ZIOP and the OP model pair is shown in columns 3 and 4 of Table 6.2. The ZIOP model is favored by all the selection criteria. The AIC and BIC tests favor the ZIOP test. For the Vuong (1989) test, the numerator and the denominator of the Vuong statistic are from the predicted values of the OP and ZIOP models respectively (see equation (6.3) for the Vuong statistic). Under the null hypothesis that the expected value of the expression in equation (6.3) is equal to 0, the Vuong statistic is less than -1.96, and this implies the ZIOP model (the denominator model) is the appropriate model.

Thus, the best of the five models that I evaluate is the ZIOP model. The ZIOP model fits the underlying smoking dataset better than the rival models. The OP model is the runner-up model in these horseraces.

6.3.2 Drinking model: selection criteria and tests of goodness of fit

Next, I conduct the horseraces for the drinking sample. I present results of these tests in Tables 6.3 and Table 6.4 below.

Table 6.3. Drinking Model Selection Criteria and Test of Goodness of Fit

	Horserace 1		Horserace 2	
	OP Model (1)	OP-ED Model (2)	OP Model (3)	SSOP Model (4)
Covariance (p-value)	-	2.451 (0.201)	-	3.458 (0.172)
ρ (p-value)	-	0.073 (0.102)	-	0.038 (0.568)
AIC	50,146.07	55,388.02	50,146.07	40,687.72
BIC	50,409.26	55,904.02	50,409.26	41,236.31

Just as in Table 6.1, columns 1 and 2 of Table 6.3 comprises horserace 1 using the drinking sample. The estimates of the covariance and correlation coefficients between the error terms of the OP and OP-ED models are 2.451 and 0.073, with p-values of 0.201 and 0.102 respectively. Under the null hypotheses that the covariance and correlation between the error terms of both models are not significantly different from 0, the OP model is the appropriate model. The AIC and BIC of the OP model are less than those of the OP-ED model, thus these results confirm that the OP is the appropriate model. From these results, I infer that marriage process in the drinking sample is random.

Next, I conduct horserace 2, the test for sample attrition (see Figure 6.1). In this case, sample attrition may cause survivorship bias if the attrition is not random. Just as in the test of specification between the OP and OP-ED models, the correlation coefficient and covariance of the error terms between the OP and Heckman sample selection ordered probit (SSOP) model pair are

the most powerful tests. These tests are shown in columns 3 and 4 of Table 6.2. In this case, I fail to reject the null hypothesis that the OP model does not suffer from sample attrition problems. But the AIC and BIC do not support this conclusion. But I consider the OP was the 'winner' of the horserace 2 because ρ and the covariance of the error terms are formal statistical test whereas the AIC and BIC provides rule of the thumb.

Next, I conduct horseraces 3 and 4 Table 6.4 below. In horserace 3, I conduct tests of fit to determine the superior model between the two inflated models – the ZIOP and ZIOPC model pair. These results are shown in columns 1 and 2 Table 6.4. Except for the BIC, all the specification tests support the conclusion that the ZIOP is the appropriate model.

Table 6.4. Drinking Model Selection Criteria and Test of Goodness of Fit

	Horserace 3		Horserace 4	
	ZIOPC Model (1)	ZIOP Model (2)	ZIOP Model (3)	OP Model (4)
Covariance (p-value)	152.298 (0.521)	-	n/a	n/a
ρ (p-value)	0.412 (0.145)	-	n/a	n/a
AIC	40,659.08	40,412.82	40,412.82	50,146.07
BIC	41,144.35	41,501.36	41,501.36	50,409.26
Vuong Test (p-value)	n/a	n/a	-10.38 (0.000)	n/a

In the final model selection, the horserace 4 is between the ZIOP and the OP model pair as I show in columns 3 and 4. Here, the ZIOP model is favored by all the selection criteria. That is, the AIC and BIC favor the ZIOP test. For the Vuong (1989) test, the numerator and the denominator of the Vuong statistic are from the predicted values of the OP and ZIOP models respectively (see equation (6.3) for the Vuong statistic). Under the null hypothesis that the expected value expression in equation (6.3) is equal to 0, the Vuong statistic is less than -1.96, and this implies the ZIOP model (the denominator model) is the appropriate model.

Thus, the best of the five models that I evaluate is the ZIOP model. The ZIOP model fits the underlying drinking dataset better than rival models. The OP model is the runner-up model in these horseraces.

Chapter 7: ESTIMATION RESULTS

In Chapter 6, I selected the overall best model after using different goodness of fit criteria to choose from the horseraces between paired models. The ZIOP model turns out to be the overall winner. As I pointed out in Chapter 6, the estimates of the OP and ZIOP models are omitted from Tables 7.1-7.6. The unconditional mean ($[Y|X]$) of an ordered probit model does not exist. So, I do not include the original estimates of these models in the result tables 7.1-7.6. Instead, I focus on the interpretation of the marginal effects of the ZIOP model in this chapter.

For comparison's sake, I show the estimates of the marginal effects of the OP model (the second runner-up model) alongside those from the ZIOP model. The marginal effects are the probabilities that particular events will occur. In this study, the events under focus are the probabilities of zero smoking and drinking "around the year of first marriage" and "after the years of first marriage". The marginal effects are shown in Tables 7.1 – 7.6.

Given the task at hand, I revisit the baseline regression of Chapter 4 in this chapter. Recall that there are four smoking and drinking ordinal outcomes (0-3) in the baseline regressions of the corresponding models in Tables 7.1-7.4. However, I focus on the interpretation of only the marginal effects of the zero smoking and drinking outcomes. After all, the ZIOP model is all about showing that inflated zeros have different data generating process (DGP) from 'genuine' and/or corner solution zeros.

Different types of marginal effects can be computed for intensities of smoking and drinking in the ZIOP model. For example, one can calculate the marginal probabilities of participation or non-participation, $pr(w = 1)$ or $pr(w = 0)$, respectively. One can also calculate the joint

marginal probabilities of zero consumption conditional on participation, $pr(y = 0, w = 1)$. Also, the unconditional marginal effects of zero tobacco and alcoholic beverage consumption, $pr(y = 0)$, can be computed under the ZIOP and OP models. All these conditional and unconditional marginal probabilities are computed and shown in Tables 7.1-7.4.

7.1 The baseline regression revisited:

The baseline regressions of smoking and drinking intensities are shown in equations (4.1) and (4.2) and of Chapter 4. These baseline regressions are spline functions, and they mimic Figure 4.1. Using the smoking intensity as an example, I present the baseline regression below:

Let Y_{it}^{0S*} be the continuous latent smoking intensity. Based on Figure 4.1, the spline specification of the impact of first marriage on smoking is:

$$\begin{aligned} Y_{it}^{0S*} &= {\beta_0}^0 + {\beta_1}^0 t + \underline{\pi^0}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}^0 && \text{if } t < 3 \\ Y_{it}^{0S*} &= {\beta_0}^1 + {\beta_1}^1 t + \underline{\pi^1}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}^1 && \text{if } 3 \le t \le 6 \\ Y_{it}^{0S*} &= {\beta_0}^2 + {\beta_1}^2 t + \underline{\pi^2}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}^2 && \text{if } t \ge 7 \end{aligned} \tag{7.1}$$

$$Y_{it}^{0S*} = \alpha_0 + \alpha_1 t + \gamma_2 d_2 + \theta_2 d_2 t + \gamma_3 d_3 + \theta_3 d_3 t + \underline{\pi}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}. \tag{7.2}$$

All the variables in equations (7.1) and (7.2) above are exactly as defined in Chapter 4. There are three marriage slopes in equation (7.1), and I use the median years as the threshold values or knots at 3 and 6 to identify the beginning of each segment of the slope. The choices of the threshold values have been explained in Chapter 4. Let the coefficients of the dummy variables d_1 , d_2, and d_3 be the intercepts or 'jumps" in Figure 4.1 during the pre-first marriage period, after first marriage period, and around first marriage period. Let $d_1 = 1$ if $t < t_1^*$, and $d_2 = 1$ if $t_1^* < t \le t_2^*$, $d_3 = 1$ if $t \ge t_2^*$, where $t_1^* = 3$ and $t_2^* = 6$. The slopes of the spline function in equation (7.2) are the parameters of the interactions of first marriage dummy variables and time. These interactions are the impact of time on the probability of smoking in cases involving single or

married smokers. Thus, the slopes of the spline function in equation (7.2) are α_1, $\alpha_1 + \theta_2$, and $\alpha_1 + \theta_2 + \theta_3$. The 'jumps' or intercepts on Figure 4.1 occur at α_0, $\alpha_0 + \gamma_2$, and $\alpha_0 + \gamma_2 + \gamma_3$.

α_1 corresponds to the slope of the baseline spline regression in the "years before first marriage", $\alpha_1 + \theta_2$ corresponds to the slope of the baseline spline regression in the "years around first marriage", and $\alpha_1 + \theta_2 + \theta_3$ is the estimate of the slope of the spline function in the "years after the first marriage". The estimates of the intercept of the baseline spline regression in equation (7.2), α_0, $\alpha_0 + \gamma_2$, and $\alpha_0 + \gamma_2 + \gamma_3$, have similar interpretation.

It should be noted that d_1 is omitted to get a full rank condition in the baseline regression in equation (7.2). The intercept parameters are also of interest in this study as they are the predicted mean amounts of tobacco product consumption (or the mean amount of alcoholic beverage consumption as the case may be).

Joining the spline function in equation (7.2) at the knots results in the specification in equations (7.3) and (7.4) below:

$$Y_{it}^{0S*} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - t_1^*) + \theta_3 d_3(t - t_2^*) + + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0S}} \quad (7.3)$$

$$Y_{it}^{0S*} = \alpha_0 + \alpha_1 t + \theta_2 d_2(t - 3) + \theta_3 d_3(t - 6) + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0S}} \quad (7.4)$$

The full derivation of equation (7.3) is shown in appendix A. The constrains $t - 3$ if $t \geq 3$ and 0 otherwise, and $t - 6$ if $t \geq 6$ and 0 otherwise apply. Plugging constraints into equation (7.3) gives equation (7.4). The coefficient of interest in equation (7.3) are α_0, α_1, θ_2, and θ_3. α_1 is the slope when $t < 3$. θ_2 is the change in the slope of the "years around first marriage". That is, θ_2 is the relative difference in the change in the probability of smoking between an individual during the "years before first marriage" and the "years around first marriage". Thus, $\alpha_1 + \theta_2$ is the slope of the "years around first marriage" segment of the spline function in Figure 4.1 and equation (7.3). θ_3 can be interpreted in a similar manner.

7.2 Marginal effects of smoking and first marriages at zero consumption levels

Tables 7.1-7.4 present the marginal effects of the OP and ZIOP models. I use these tables to answer the research questions and hypothesis posed in Chapters 1 and 4 respectively. For example, what does first marriage tell us about intensities of drinking and smoking? What are impacts of socio-economic and demographics on the probabilities of smoking and drinking? To establish the validity of hypotheses 1 and 2, I show separate tables for the overall sample, and for male and female subsamples of smoking intensities in Tables 7.1 and 7.2 respectively. For hypothesis 3 or 4, for drinking sample, I also present the marginal effects of the overall sample, and by gender in Tables 7.3 and 7.4, respectively.

Table 7.1 below presents the marginal effects of zero tobacco consumption. Unlike the conditional marginal effect of zero smoking intensity in Table 7.3 (column 3), the estimates of the consumption decision conditional on participation $pr(y = 0, w = 1)$ are larger and significantly different from zero for some of the predictors. These estimates not only imply that $pr(y = 0, w = 1)$ contributes to the unconditional marginal effects of zero tobacco consumption $pr(y = 0)$ under the ZIOP model, they also imply that the marginal effects of the OP and the ZIOP models of smoking intensities will be significantly different for most predictors. In practical terms, the implication of this result is that there are relatively more infrequent smokers in the smoking sample than there are infrequent drinkers in the drinking sample (see Table 7.3). A cursory glance at Tables 3.3 and 3.4 in section 3.2 confirms this result.

Table 7.1. Smoking: Marginal Effects of Zero Consumption

		ZIOP		
VARIABLES	(1) OP (y=0)	(2) pr(w=0)	(3) pr(y=0, w=1)	(4) pr(y=0)
Years around first marriage (Intercept)	-0.0268	0.0367	-0.0571*	-0.0204
	(0.0206)	(0.0401)	(0.0330)	(0.0209)
Years after first marriage (Intercept)	0.0237	0.0961**	-0.0640	0.0322
	(0.0255)	(0.0488)	(0.0399)	(0.0259)
Years before first marriage (Slope)	-0.0123	0.0104	-0.0203	-0.00992
	(0.0120)	(0.0233)	(0.0192)	(0.0121)
Years around first marriage (Slope)	0.0199***	0.0179*	0.00240	0.0203***
	(0.00526)	(0.00959)	(0.00780)	(0.00534)
Years after first marriage (Slope)	0.0157***	0.0187***	-0.00209	0.0166***
	(0.00300)	(0.00545)	(0.00448)	(0.00304)
Age	-0.110***	-0.0786***	-0.0360**	-0.115***
	(0.0122)	(0.0221)	(0.0165)	(0.0122)
High School or Less	0.159***	0.139***	0.0312***	0.170***
	(0.00872)	(0.0136)	(0.00952)	(0.00944)
Bachelor's Degree	0.331***	0.219***	0.117***	0.337***
	(0.0103)	(0.0218)	(0.0187)	(0.0109)
Hispanics	-7.99e-05	-0.481***	0.467***	-0.0143*
	(0.00727)	(0.0493)	(0.0466)	(0.00790)
Non-Black, Non-Hispanics	-0.180***	-0.0222	-0.145***	-0.167***
	(0.00664)	(0.0415)	(0.0391)	(0.00696)
Female	0.0482***	-0.0294**	0.0714***	0.0420***
	(0.00533)	(0.0120)	(0.0104)	(0.00539)
Single, Non-Cohabiting	0.109***	0.0185	0.0773*	0.0958**
	(0.0312)	(0.0692)	(0.0415)	(0.0401)
Married	0.0711**	-0.0243	0.0803*	0.0560
	(0.0335)	(0.0715)	(0.0443)	(0.0417)
Separating	-0.108***	-0.184***	0.0526	-0.132***
	(0.0259)	(0.0530)	(0.0349)	(0.0318)
Out-of-Labor Force	0.0993***	0.0943***	0.00758	0.102***
	(0.0119)	(0.0206)	(0.0156)	(0.0124)
Employed	0.0692***	0.0352*	0.0336**	0.0688***
	(0.0109)	(0.0189)	(0.0144)	(0.0114)
Self-Reported Health – very good	-0.0816***	-0.0652***	-0.0139	-0.0791***
	(0.00612)	(0.0117)	(0.00992)	(0.00612)
Self-Reported Health - Good	-0.128***	-0.146***	0.0118	-0.134***
	(0.00685)	(0.0132)	(0.0110)	(0.00698)
Self-Reported Health - Fair	-0.176***	-0.153***	-0.0261*	-0.179***
	(0.0115)	(0.0202)	(0.0153)	(0.0123)
Cohabit	0.0340	-0.0510	0.0786	0.0276
	(0.0271)	(0.0664)	(0.0505)	(0.0306)
Insurance	0.0913***	0.0749***	0.0163*	0.0912***
	(0.00607)	(0.0112)	(0.00872)	(0.00622)
ln(Wage)	0.00137**	0.00167*	-0.000645*	0.00103*
	(0.000631)	(0.000985)	(0.000381)	(0.000606)
ln(CPI Drinks)	0.256*	0.403*	-0.155*	0.247*
	(0.137)	(0.210)	(0.0820)	(0.129)
ln(Marijuana)	-0.182***	-0.265***	0.102***	-0.163***
	(0.0365)	(0.0556)	(0.0227)	(0.0339)
ln(CPI Cigarette)	-0.0260	-0.0459	0.0177	-0.0282
	(0.0421)	(0.0641)	(0.0247)	(0.0394)
Observations	27,031	27,031	27,031	27,031

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

In terms of policy implications, one might be interested in the pre-and post-marriage magnitude of the smoking intensity's $pr(y = 0, w = 1)$, and this probability is needed for some analyses in this section.

Also, in Table 7.1 an average respondent is more likely to be a smokers during the "years around first marriage" than during the "years before first marriage" under the ZIOP (OP) model. Specifically, there is a 2.04 % (2.68%) (see columns 4 and 1 respectively) chance of a *decrease* in the predicted mean of zero tobacco consumption between the two periods. Since the predicted mean of smoking rises during the "years around first marriage", these estimates are counterintuitive. But the estimates and their difference are not significantly different from zero.

For the "years after first marriage", the probability of the predicated mean of zero tobacco consumption increases by 3.22% (2.37%) (see columns 4 and 1 respectively) according to the ZIOP (OP) model. As a standalone marginal effect, the probability of predicted mean of nonparticipation, $pr(w = 0)$, increases by 9.61% (see column 2) for this outcome, and this estimate is significant. This estimate confirms what we already know from Table 3.3 in Chapter 3 – most nonsmokers are lifelong abstainers, and first marriages reinforce their perfectly inelastic demand for tobacco products. Another take away from Table 7.1 is that the "after first marriage" intercept under the ZIOP model, $pr(w = 0) = 0.0961$, is larger than $pr(y = 0) = 0.0322$, and this result means that most of the respondents are non-smokers.

The slopes of the spline function "before the year of first marriage" are not significantly different from zero. For the "years around first marriage", the ZIOP model (the OP model) predicts a 2.03% (1.99%) (see columns 4 and 1 respectively) increase in the probability of zero tobacco consumption. Finally, the ZIOP model (the OP model) implies that "after the year of first marriage the probability of observing zero tobacco consumption is 1.66% (1.57%). The last set of results are also significantly different from zero. Overall, a first marriage encourages non-smoking

behavior. The estimate of the slope "around the years of first marriage" and "after the year of first marriage" confirm **hypothesis 1** in section 4.2 of this study.

The superiority of the ZIOP over the OP model is evident in Table 7.1. The marginal effects of the ZIOP model are bigger than those under the OP model. Thus, using the OP model instead of the ZIOP model can lead to wrong policy prescriptions on smoking reduction.

The marginal effects of other variables in Table 7.1 have similar interpretations. For example, the probabilities of tobacco consumption fall at all the levels educational attainment relative the respondents with no education (the reference variable). On average, females smoke more than males (the reference variable). An increase in wages increases the probability of zero tobacco consumption. Marijuana and tobacco are complements, while alcohol beverage and tobacco are substitutes.

Separating the smoking sample by gender in Table 7.2 panels A and B below, the unconditional probability of observing zero alcohol consumption for a male (female) "around the years of first marriage" is 2.12% (2.39%). For a male (a female), the probability of zero consumption of alcohol "after the year of first marriage" increased by 2.22% (1.44%). Combined, these two results confirm **hypothesis 2**.

7.3 Marginal effects of drinking and first marriages at the zero consumption level

Under the ZIOP and OP models in Table 7.3 below, relative to the "years before first marriage", married respondents are more likely to be non-drinkers in the "years around first marriage". Specifically, there is a 2.84% (2.55%) increase in the probability of the predicted mean (the intercept of the spline function) of zero alcohol consumption between "the years around first marriage" and the "years before first marriage" under the ZIOP (OP) model.

Table 7.2. Smoking: Marginal Effects of Zero Consumption (Male and Female)

	Male		Female	
	OP	ZIOP	OP	ZIOP
Variables	Pr(y=0)	pr(y=0)	Pr(y=0)	pr(y=0)
Years around First Marriage (Intercept)	-0.0429	-0.0376	-0.0135	1.135
	(0.0307)	(0.0308)	(0.0278)	(25.72)
Years after First Marriage (Intercept)	0.0150	0.0185	0.0291	1.185
	(0.0386)	(0.0389)	(0.0341)	(25.72)
Years before First Marriage (Slope)	-0.0221	-0.0217	-0.00425	0.513
	(0.0178)	(0.0180)	(0.0162)	(12.86)
Year before First Marriage (Slope)	0.0209***	0.0212***	0.0184***	0.0239***
	(0.00801)	(0.00812)	(0.00696)	(0.00841)
First Marriage Slope (After)	0.0196***	0.0222***	0.0127***	0.0114***
	(0.00470)	(0.00476)	(0.00388)	(0.00424)
Age	-0.114***	-0.125***	-0.0994***	-0.0707***
	(0.0188)	(0.0187)	(0.0162)	(0.0199)
High School or Less	0.178***	0.203***	0.140***	0.152***
	(0.0126)	(0.0137)	(0.0121)	(0.0137)
Bachelor's Degree	0.338***	0.357***	0.320***	0.330***
	(0.0155)	(0.0162)	(0.0140)	(0.0151)
Hispanics	0.0173	-0.00421	-0.0174*	-0.0181*
	(0.0114)	(0.0124)	(0.00927)	(0.00929)
Non-Black, Non-Hispanics	-0.140***	-0.117***	-0.216***	-0.215***
	(0.0102)	(0.0108)	(0.00864)	(0.00867)
Female	-	-	-	-
Single, Non-Cohabiting	0.213***	0.207***	0.0539	0.0802
	(0.0521)	(0.0628)	(0.0373)	(0.0555)
Married	0.175***	0.171***	0.0228	0.0498
	(0.0552)	(0.0656)	(0.0406)	(0.0574)
Separating	0.107***	0.111***	0.0953***	0.111***
	(0.0185)	(0.0192)	(0.0159)	(0.0174)
Out-of-Labor Force	0.0657***	0.0679***	0.0734***	0.0917***
	(0.0159)	(0.0166)	(0.0152)	(0.0167)
Employed	-0.0861***	-0.0809***	-0.0761***	-0.0780***
	(0.00911)	(0.00908)	(0.00829)	(0.00833)
Self-Reported Health – very good	-0.130***	-0.135***	-0.126***	-0.122***
	(0.0106)	(0.0107)	(0.00904)	(0.00910)
Self-Reported Health - Good	-0.181***	-0.187***	-0.174***	-0.170***
	(0.0185)	(0.0199)	(0.0148)	(0.0150)
Self-Reported Health - Fair	-0.128**	-0.160**	-0.259***	-0.254***
	(0.0572)	(0.0623)	(0.0388)	(0.0393)
Cohabit	0.121***	0.121***	-0.0116	0.0817
	(0.0399)	(0.0448)	(0.0351)	(0.0709)
Insurance	0.0965***	0.0963***	0.0858***	0.0801***
	(0.00945)	(0.00962)	(0.00870)	(0.00839)
ln(Wage)	0.00267***	0.00141	0.000293	-0.000118
	(0.000995)	(0.000963)	(0.000814)	(0.000738)
ln(CPI Drinks)	0.00243	0.111	0.445**	0.161
	(0.211)	(0.198)	(0.180)	(0.180)
ln(Marijuana)	-0.215***	-0.210***		
	(0.0557)	(0.0519)		
ln(CPI Cigarette)	-0.0161	-0.0190	-0.0254	0.0122
	(0.0638)	(0.0601)	(0.0559)	(0.0549)
Observations	12,250	12,250	14,781	14,781

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table 7.3. Drinking: Marginal Effects of Zero Consumption

	(1) OP	(2) ZIOP	(3) ZIOP	(4) ZIOP
Variables	Pr(y=0)	pr(w=0)	pr(y=0, w=1)	p(y=0)
Years around first marriage (intercept)	0.0255	0.0284	-1.53e-05	0.0284
	(0.0204)	(0.0231)	(0.000124)	(0.0231)
Years after first marriage (intercept)	0.118***	0.132***	8.72e-05	0.132***
	(0.0252)	(0.0284)	(0.000414)	(0.0283)
Years before first marriage (slope)	0.00562	0.00695	-1.30e-05	0.00694
	(0.0119)	(0.0135)	(8.36e-05)	(0.0135)
Years around first marriage (slope)	0.0301***	0.0332***	2.91e-05	0.0333***
	(0.00524)	(0.00594)	(0.000136)	(0.00593)
Years after first marriage (slope)	0.0113***	0.0149***	1.02e-06	0.0149***
	(0.00288)	(0.00316)	(1.57e-05)	(0.00316)
Age	-0.118***	-0.118***	-6.34e-06	-0.118***
	(0.0104)	(0.0115)	(4.72e-05)	(0.0115)
High School or Less	-0.0297***	-0.0686***	0.000105	-0.0685***
	(0.00848)	(0.00945)	(0.000574)	(0.00946)
Bachelor's Degree	-0.0315***	-0.118***	0.000478	-0.117***
	(0.0108)	(0.0125)	(0.00244)	(0.0122)
Hispanics	-0.155***	-0.120***	-0.000599	-0.121***
	(0.00863)	(0.00958)	(0.00275)	(0.00935)
Non-Black, Non-Hispanics	-0.191***	-0.193***	-0.000459	-0.193***
	(0.00754)	(0.00828)	(0.00204)	(0.00808)
Female	0.119***	0.0754***	0.000435	0.0758***
	(0.00520)	(0.00626)	(0.00216)	(0.00584)
Single, Non-Cohabiting	0.110***	0.0794**	0.000316	0.0797**
	(0.0287)	(0.0335)	(0.00153)	(0.0334)
Married	0.105***	0.0628*	0.000429	0.0633*
	(0.0305)	(0.0355)	(0.00205)	(0.0354)
Separating	-0.0143	-0.0179	1.40e-05	-0.0179
	(0.0221)	(0.0258)	(8.38e-05)	(0.0258)
Out-of-Labor Force	0.0720***	0.0703***	4.14e-05	0.0703***
	(0.0122)	(0.0133)	(0.000193)	(0.0133)
Employed	-0.00128	-0.00967	4.84e-05	-0.00962
	(0.0108)	(0.0120)	(0.000240)	(0.0120)
Self-Reported Health – very good	-0.0453***	-0.0588***	5.33e-05	-0.0587***
	(0.00616)	(0.00691)	(0.000268)	(0.00690)
Self-Reported Health - Good	-0.0448***	-0.0394***	-8.70e-05	-0.0394***
	(0.00680)	(0.00759)	(0.000411)	(0.00758)
Self-Reported Health - Fair	-0.0632***	-0.0517***	-0.000135	-0.0518***
	(0.0107)	(0.0119)	(0.000650)	(0.0119)
Cohabit	0.0868***	0.0243	0.00120	0.0255
	(0.0308)	(0.0339)	(0.00517)	(0.0340)
Insurance	0.00279	-0.00920	7.85e-05	-0.00912
	(0.00622)	(0.00691)	(0.000377)	(0.00691)
ln(Wage)	-0.00465***	-0.00505***	2.54e-06	-0.00504***
	(0.000629)	(0.000689)	(1.31e-05)	(0.000688)
ln(CPI Drinks)	0.228*	-0.0372	1.87e-05	-0.0372
	(0.136)	(0.152)	(0.000122)	(0.151)
ln(Marijuana)	-0.0923***	-0.0517	2.60e-05	-0.0517
	(0.0337)	(0.0374)	(0.000135)	(0.0374)
ln(CPI Cigarette)	-0.00228	0.0443	-2.23e-05	0.0443
	(0.0406)	(0.0454)	(0.000117)	(0.0454)
Observations	27,583	27,583	27,583	27,583

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Thus, the predicted mean of zero alcohol consumption falls during the "years around first marriage". But these estimates are not significant under both OP and ZIOP models.

The biggest increase in predicted mean (the intercept) of non-drinking outcome occurs during the "years after of first marriage". During the "years after first marriage", the probability of unconditional predicted mean of zero alcohol consumption increases by 13.2% (11.8%) according to the ZIOP model (OP model). In this case, the estimate is statistically significant significant. Also, the joint probability of zero consumption conditional on participation, $pr(y = 0, w = 1) = 0.0000872$. This estimate contributes very little to the unconditional probability of the predicted mean of zero alcohol consumption, $pr(y = 0) = 0.132$, in column 4 of Table 7.1. Indeed, none of the estimates in the column 3 of Table 7.3 are significantly different from zero, and the probability of non-participation, $pr(w = 0)$ is approximately equal to the unconditional probability of zero alcohol consumption, $pr(y = 0)$, under the ZIOP model. These results confirm the alcohol consumption univariate analysis in Table 3.4, section 3.2.

Turning to the slope segments of the spline regression, the ZIOP model (OP model) predicts an increase of 0.694% (0.562%) in the probability of zero alcohol consumption during the "years before first marriage". However, these predicted probabilities are not significantly different from zero. The biggest impact of first marriage on zero alcohol consumption occurs during the "years around first marriage". In this case, the ZIOP model (OP model) predicts a 3.33% (3.01%) increase in the probability of zero consumption of alcohol. And just like the intercepts, the probability of nonparticipation, $pr(w = 0)$, is driving the unconditional marginal effects of the ZIOP model, $pr(y = 0)$. Finally, for the "year after first marriage", the ZIOP model (OP model) implies that the probability of observing a zero consumption of alcohol is 1.49% (1.13%).

Overall, a first marriage encourages non-drinking behavior. The slope "around the years of first marriage" and "after the year of first marriage" confirm **hypothesis 3** of this study. This

benefit in reduced drinking that flows from first marriage gets bigger as a respondent gets closer to the year of first marriage, and the benefit continues after the year of first marriage. Despite the minuscule contribution of $pr(y = 0, w = 1)$, the overall superiority of the ZIOP over the OP model is evident in Table 7.3. For example, using the OP model instead of the ZIOP model for policy understates the probability of observing the zero outcome of alcohol consumption. Since the ZIOP model is the appropriate model, wrong prescriptions may be adopted by policy makers if such prescriptions are based on the ZIOP estimates. A policy mistake can result in misallocation of financial and human resources.

The marginal effects of other variables in Table 7.3 have similar interpretations like the slope and intercept of the spline function. For example, the probabilities of alcohol consumption increase at all the educational attainment levels relative to little or no educational attainment level (the reference variable). On average, females drink more than males (the reference variable). Individuals who self-report "excellent" health status (the reference variable) drink less than other categories of self-reported health status. A 10% increase in wages increases the probability of zero alcohol consumption by 5.04%, making tobacco a normal good. The unconditional marginal effect is negatively responsive to own-price, as a 10% increase in own price implies a 37.2% reduction zero alcohol consumption. Marijuana and alcoholic beverage are complements.

To establish the validity of **hypothesis 4** in section 4.1, I show separate marginal effect estimates for male and female subsamples in Table 7.4 below. The male and female marginal effects are respectively shown in column 1 and 2, and 3 and 4. In both panels, the columns for the joint probabilities of consumption conditional on participation, $pr(y = 0, w = 1)$, are omitted because the estimates of all the marginal effects are statistically zeros (see Table 7.3). To validate hypothesis 4, I focus on the slope of the baseline regressions in Table 7.4.

Table 7.4. Drinking: Marginal Effects of Zero Consumption (Male and Female)

	Male			Female		
	OP	ZIOP	ZIOP	OP	ZIOP	ZIOP
Variables	Pr(y=0)	pr(w=0)	p(y=0)	Pr(y=0)	pr(w=0)	p(y=0)
Years around First Marriage (Intercept)	0.0152	0.0119	0.0119	0.0356	0.0455	0.0455
	(0.0278)	(0.0318)	(0.0318)	(0.0301)	(0.0329)	(0.0329)
Years after First Marriage (Intercept)	0.0826**	0.0778*	0.0778*	0.148***	0.177***	0.177***
	(0.0349)	(0.0402)	(0.0402)	(0.0367)	(0.0396)	(0.0396)
Years before First Marriage (Slope)	-0.00407	-0.00782	-0.00782	0.0167	0.0221	0.0221
	(0.0162)	(0.0186)	(0.0186)	(0.0176)	(0.0192)	(0.0192)
Year before First Marriage (Slope)	0.0206***	0.0196**	0.0196**	0.0388***	0.0439***	0.0439***
	(0.00731)	(0.00851)	(0.00851)	(0.00754)	(0.00820)	(0.00820)
First Marriage Slope (After)	0.0108***	0.0137***	0.0137***	0.0117***	0.0153***	0.0153***
	(0.00416)	(0.00479)	(0.00479)	(0.00403)	(0.00421)	(0.00421)
Age	-0.125***	-0.111***	-0.111***	-0.109***	-0.131***	-0.131***
	(0.0148)	(0.0167)	(0.0167)	(0.0150)	(0.0158)	(0.0158)
High School or Less	0.00699	-0.0426***	-0.0426***	-0.0779***	-0.0942***	-0.0942***
	(0.0112)	(0.0133)	(0.0133)	(0.0128)	(0.0133)	(0.0133)
Bachelor's Degree	0.0130	-0.101***	-0.101***	-0.0898***	-0.139***	-0.139
	(0.0146)	(0.0172)	(0.0172)	(0.0160)	(0.0170)	(2.905)
Hispanics	-0.00792	-0.177***	-0.177***	-0.119***	-0.199***	-0.199***
	(0.0250)	(0.0278)	(0.0278)	(0.0245)	(0.0266)	(0.0266)
Non-Black, Non-Hispanics	-0.180***	-0.133***	-0.133***	-0.127***	-0.110***	-0.110***
	(0.0122)	(0.0135)	(0.0135)	(0.0123)	(0.0127)	(0.0127)
Female	-	-	-	-	-	-
	-	-	-	-	-	-
Single, Non-Cohabiting	0.0875**	0.0204	0.0204	0.141***	0.117***	0.117***
	(0.0433)	(0.0552)	(0.0575)	(0.0386)	(0.0427)	(0.0427)
Married	0.0782*	-0.0104	-0.0104	0.145***	0.116**	0.116
	(0.0454)	(0.0574)	(0.0596)	(0.0414)	(0.0458)	(3.140)
Separating	0.00600	-0.0144	-0.0144	-0.0246	-0.0191	-0.0191
	(0.0324)	(0.0425)	(0.0454)	(0.0304)	(0.0335)	(0.0335)
Out-of-Labor Force	0.0634***	0.0655***	0.0655***	0.0737***	0.0665***	0.0665
	(0.0179)	(0.0200)	(0.0200)	(0.0172)	(0.0182)	(1.092)
Employed	0.000282	-0.00305	-0.00305	-0.000307	-0.00897	-0.00897
	(0.0145)	(0.0164)	(0.0164)	(0.0161)	(0.0171)	(1.092)
Self-Reported Health – very good	-0.0496***	-0.0662***	-0.0662***	-0.0386***	-0.0481***	-0.0481
	(0.00840)	(0.00970)	(0.00970)	(0.00909)	(0.00972)	(2.347)
Self-Reported Health - Good	-0.0493***	-0.0440***	-0.0440***	-0.0392***	-0.0310***	-0.0310
	(0.00956)	(0.0111)	(0.0111)	(0.00978)	(0.0104)	(0.311)
Self-Reported Health - Fair	-0.0489***	-0.0269	-0.0269	-0.0730***	-0.0623***	-0.0623
	(0.0160)	(0.0186)	(0.0186)	(0.0149)	(0.0158)	(4.570)
Cohabit	0.0625	-0.0520	-0.0520	0.118***	0.0808*	0.0808*
	(0.0483)	(0.0501)	(0.0501)	(0.0396)	(0.0431)	(0.0444)
Insurance	-0.0224***	-0.0485***	-0.0485***	0.0352***	0.0320***	0.0320***
	(0.00848)	(0.00961)	(0.00962)	(0.00922)	(0.00986)	(0.00986)
ln(Wage)	-.00445***	-0.00461***	-0.00461***	-0.00417***	-0.00480***	-0.00480***
	(0.000919)	(0.00103)	(0.00103)	(0.000878)	(0.000926)	(0.000926)
ln(CPI Drinks)	0.482**	0.0757	0.0757	-0.0438	-0.146	-0.146
	(0.192)	(0.221)	(0.221)	(0.194)	(0.207)	(0.207)
ln(Marijuana)	-0.0630	0.0223	0.0223	-0.135***	-0.137***	-0.137***
	(0.0472)	(0.0539)	(0.0539)	(0.0486)	(0.0516)	(0.0516)
ln(CPI Cigarette)	-0.0492	0.0124	0.0124	0.0594	0.0837	0.0837
	(0.0567)	(0.0655)	(0.0655)	(0.0585)	(0.0623)	(0.0623)
Observations	12,408	12,408	12,408	15,175	15,175	15,175

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7.4, the unconditional probability of observing zero alcohol consumption for a male (female) "around the years of first marriage" is 1.96% (4.39%). For a male (a female), the probability of zero consumption of alcohol "after the year of first marriage" increased by 1.37% (1.53%). Combined, these two results confirm **hypothesis 4**.

7.4 The marginal effects of alcoholic beverage and tobacco consumptions at other consumption levels

I present the marginal effects of tobacco and alcoholic beverage consumptions at other ordinal outcome levels in Tables 7.5 and 7.6, respectively. These tables corroborate the results in Tables 7.1 and 7.3, albeit at different level of ordinal outcomes.

Turning my attention to Table 7.5, the during the "years around first marriage" under the ZIOP model, the probability of smoking falls by 0.616%, 1.33%, 0.0827% at the first, the second, and the third tobacco consumption intensity levels respectively. These results confirm that the benefits of first marriage in terms of reduced smoking is pervasive at all levels during the "years around first marriage".

Similarly, during the "years after the first marriage" under the ZIOP model in Table 7.5, the probability of smoking falls by 0.575%, 1.03%, and 0.0554% at the first, the second, and the third tobacco consumption intensity levels respectively. These estimates also confirm that the benefit of first marriage in terms of reduced smoking occurs at all levels.

Another take away from the estimates of the marriage slope in Table 7.5 is that the benefits of marriage on smoking is diminishing in the level of intensity of tobacco consumption. These benefits peaked at second consumption level. The benefits fall thereafter, and it is almost completely gone at the third consumption level.

Table7.5. Tobacco Consumption: Marginal Effects of Non-Zero Consumption Levels

Variables	OP Pr(y=1)	ZIOP Pr(y=1)	OP Pr(y=2)	ZIOP Pr(y=2)	OP Pr(y=3)	ZIOP Pr(y=3)
Around Marriage (intercept)	0.00757	-0.00330	0.0181	0.0212	0.00118	0.00251**
	(0.00582)	(0.00909)	(0.0139)	(0.0139)	(0.000907)	(0.00124)
After Marriage (intercept)	-0.00668	-0.0215*	-0.0160	-0.0114	-0.00104	0.000766
	(0.00721)	(0.0111)	(0.0172)	(0.0172)	(0.00112)	(0.00151)
Before Marriage (Slope)	0.00346	-0.000308	0.00826	0.00924	0.000537	0.000990
	(0.00339)	(0.00528)	(0.00809)	(0.00807)	(0.000527)	(0.000722)
Around Marriage (Slope)	-0.00561***	-0.00616***	-0.0134***	-0.0133***	-0.000870***	-0.000827***
	(0.00149)	(0.00223)	(0.00355)	(0.00356)	(0.000239)	(0.000310)
After Marriage (Slope)	-0.00444***	-0.00575***	-0.0106***	-0.0103***	-0.000688***	-0.000554***
	(0.000850)	(0.00126)	(0.00203)	(0.00204)	(0.000141)	(0.000181)
Age	0.0310***	0.0308***	0.0740***	0.0784***	0.00481***	0.00535***
	(0.00348)	(0.00526)	(0.00825)	(0.00778)	(0.000637)	(0.000720)
High School or Less	-0.0936***	-0.0911***	-0.223***	-0.231***	-0.0141***	-0.0146***
	(0.00300)	(0.00478)	(0.00762)	(0.00776)	(0.00121)	(0.00132)
Bachelor's Degree	-0.131***	-0.125***	-0.261***	-0.277***	-0.0147***	-0.0154***
	(0.00632)	(0.00775)	(0.00858)	(0.00841)	(0.00125)	(0.00136)
Hispanics	3.12e-05	0.0344***	4.71e-05	-0.0193***	1.60e-06	-0.000756***
	(0.00284)	(0.00722)	(0.00429)	(0.00525)	(0.000146)	(0.000188)
Non-Black, Non-Hispanics	0.0545***	0.0198***	0.119***	0.138***	0.00680***	0.00911***
	(0.00247)	(0.00534)	(0.00427)	(0.00484)	(0.000530)	(0.000723)
Female	-0.0136***	-0.000949	-0.0325***	-0.0373***	-0.00211***	-0.00375***
	(0.00151)	(0.00230)	(0.00362)	(0.00365)	(0.000277)	(0.000406)
Single, Non-Cohabiting	-0.0295***	-0.0145	-0.0745***	-0.0741***	-0.00496***	-0.00718***
	(0.00736)	(0.0169)	(0.0222)	(0.0248)	(0.00180)	(0.00249)
Married	-0.0180**	-0.000489	-0.0495**	-0.0496*	-0.00358*	-0.00598**
	(0.00780)	(0.0175)	(0.0238)	(0.0261)	(0.00192)	(0.00263)
Separating	0.0178***	0.0475***	0.0812***	0.0799***	0.00897***	0.00416
	(0.00533)	(0.0134)	(0.0192)	(0.0213)	(0.00226)	(0.00266)
Out-of-Labor Force	-0.0261***	-0.0294***	-0.0683***	-0.0678***	-0.00488***	-0.00469***
	(0.00291)	(0.00508)	(0.00841)	(0.00846)	(0.000760)	(0.000949)
Employed	-0.0172***	-0.0142***	-0.0483***	-0.0503***	-0.00368***	-0.00426***
	(0.00244)	(0.00463)	(0.00785)	(0.00790)	(0.000717)	(0.000913)
Health – very Good	0.0263***	0.0250***	0.0527***	0.0513***	0.00260***	0.00281***
	(0.00205)	(0.00261)	(0.00395)	(0.00402)	(0.000277)	(0.000358)
Health – Good	0.0384***	0.0470***	0.0848***	0.0827***	0.00475***	0.00404***
	(0.00215)	(0.00305)	(0.00461)	(0.00464)	(0.000438)	(0.000458)
Health - Fair	0.0487***	0.0522***	0.120***	0.119***	0.00767***	0.00775***
	(0.00274)	(0.00508)	(0.00837)	(0.00843)	(0.000886)	(0.00103)
Cohabit	-0.00993	0.00276	-0.0226	-0.0274	-0.00144	-0.00294**
	(0.00817)	(0.0144)	(0.0178)	(0.0179)	(0.00112)	(0.00123)
Insurance	-0.0258***	-0.0267***	-0.0615***	-0.0606***	-0.00400***	-0.00388***
	(0.00176)	(0.00265)	(0.00412)	(0.00409)	(0.000392)	(0.000436)
ln(Wage)	-0.000388**	-0.000444*	-0.000926**	-0.000564*	-6.02e-05**	-1.83e-05*
	(0.000178)	(0.000262)	(0.000425)	(0.000333)	(2.80e-05)	(1.09e-05)
ln(CPI Drinks)	-0.0724*	-0.107*	-0.173*	-0.136*	-0.0112*	-0.00442*
	(0.0388)	(0.0558)	(0.0925)	(0.0708)	(0.00607)	(0.00233)
ln(Marijuana)	0.0513***	0.0703***	0.122***	0.0894***	0.00796***	0.00290***
	(0.0103)	(0.0147)	(0.0246)	(0.0187)	(0.00170)	(0.000648)
ln(CPI Cigarette)	0.00735	0.0122	0.0175	0.0155	0.00114	0.000503
	(0.0119)	(0.0170)	(0.0284)	(0.0216)	(0.00185)	(0.000704)
Observations	27,031	27,031	27,031	27,031	27,031	27,031

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Similarly, during the "years after the first marriage" under the ZIOP model in Table 7.6, the probability of drinking falls by 1.12%, 0.0352%, and 0.0159% at the first, the second, and the third alcohol consumption intensity levels respectively. These estimates confirm that the benefits of first marriage in terms of reduced drinking occur at all levels during the "year around first marriage" and the "year after first marriage" Table 7.6 shows that during the "years around first marriage" under the ZIOP model, the probability of drinking falls by 2.14%, 1.13%, and 0.059% at the first, the second, and the third alcohol consumption intensity levels respectively.

Another take away from Table 7.6 is that the benefits of marriage on drinking is diminishing in level of intensity of alcohol consumption. That is, this benefit peaked at the first consumption intensity level. The benefits reduced thereafter, and it is almost completely gone at the third consumption level.

Contrasting Tables 7.1, 7.3, 7.5 and 7.6, one will reach the same conclusion that the benefits of marriage on smoking and drinking is pervasive at all consumption intensity levels. However, the benefits of marriage on drinking peaked earlier than the benefits of marriage on smoking. The reasons for these different peaks on the spline functions of smoking and drinking can be traced to the consumption behavior of the respondents with less than perfectly inelastic demand during the 30-day survey period. Since the previous estimates in Tables 7.1 and 7.3, and univariate analysis in Tables 3.3 and 3.4 have shown that there are more infrequent smokers than infrequent drinkers, it is expected that the benefits of marriage last longer for smokers.

Table 7.6. Drinking: Marginal Effects of Non-Zero Consumption Levels

Variables	OP Pr(y=1)	ZIOP Pr(y=1)	OP Pr(y=2)	ZIOP Pr(y=2)	OP Pr(y=3)	ZIOP Pr(y=3)
Marriage Intercept (Around)	-0.0132	-0.0238	-0.0115	-0.00450	-0.000783	-0.000153
	(0.0105)	(0.0230)	(0.00924)	(0.0126)	(0.000631)	(0.000827)
Marriage Intercept (After)	-0.0608***	-0.0889***	-0.0533***	-0.0413**	-0.00362***	-0.00210*
	(0.0130)	(0.0288)	(0.0114)	(0.0162)	(0.000840)	(0.00109)
First Marriage Slope (Before)	-0.00290	-0.00707	-0.00254	8.89e-05	-0.000173	4.34e-05
	(0.00614)	(0.0135)	(0.00539)	(0.00732)	(0.000366)	(0.000481)
First Marriage Slope (Around)	-0.0155***	-0.0214***	-0.0136***	-0.0113***	-0.000924***	-0.000590**
	(0.00271)	(0.00602)	(0.00238)	(0.00339)	(0.000181)	(0.000230)
First Marriage Slope (After)	-0.00584***	-0.0112***	-0.00513***	-0.00352*	-0.000348***	-0.000159
	(0.00149)	(0.00338)	(0.00131)	(0.00204)	(9.40e-05)	(0.000137)
Age	0.0610***	0.0894***	0.0535***	0.0277***	0.00363***	0.00125***
	(0.00541)	(0.0115)	(0.00478)	(0.00631)	(0.000454)	(0.000428)
High School or Less	0.0170***	0.189***	0.0136***	-0.0668***	0.000874***	-0.00502***
	(0.00592)	(0.0126)	(0.00460)	(0.00745)	(0.000301)	(0.000716)
Bachelor's Degree	0.0295***	0.275***	0.0262***	-0.0879***	0.00178***	-0.00573***
	(0.00836)	(0.0213)	(0.00838)	(0.0106)	(0.000633)	(0.000766)
Hispanics	0.100***	0.0366***	0.0521***	0.0801***	0.00256***	0.00428***
	(0.00584)	(0.00936)	(0.00298)	(0.00475)	(0.000287)	(0.000493)
Non-Black, Non-Hispanics	0.118***	0.123***	0.0697***	0.0677***	0.00377***	0.00282***
	(0.00540)	(0.00807)	(0.00243)	(0.00363)	(0.000364)	(0.000320)
Female	-0.0610***	0.0226***	-0.0540***	-0.0932***	-0.00349***	-0.00515***
	(0.00279)	(0.00601)	(0.00250)	(0.00364)	(0.000343)	(0.000519)
Single, Non-Cohabiting	-0.0485***	0.0209	-0.0574***	-0.0921***	-0.00458***	-0.00849**
	(0.00996)	(0.0373)	(0.0174)	(0.0261)	(0.00172)	(0.00368)
Married	-0.0455***	0.0438	-0.0552***	-0.0982***	-0.00444**	-0.00894**
	(0.0106)	(0.0393)	(0.0183)	(0.0274)	(0.00177)	(0.00374)
Separating	0.00381	0.0199	0.00950	-0.00144	0.000973	-0.000568
	(0.00624)	(0.0313)	(0.0144)	(0.0232)	(0.00146)	(0.00317)
Out-of-Labor Force	-0.0409***	-0.0475***	-0.0294***	-0.0217***	-0.00176***	-0.00110**
	(0.00659)	(0.0136)	(0.00533)	(0.00794)	(0.000382)	(0.000554)
Employed	0.000639	0.0146	0.000602	-0.00462	4.13e-05	-0.000382
	(0.00539)	(0.0123)	(0.00506)	(0.00721)	(0.000347)	(0.000508)
Health – very good	0.0243***	0.0519***	0.0197***	0.00671*	0.00128***	0.000149
	(0.00337)	(0.00705)	(0.00266)	(0.00390)	(0.000206)	(0.000238)
Health - Good	0.0241***	0.0165**	0.0194***	0.0217***	0.00126***	0.00128***
	(0.00366)	(0.00779)	(0.00298)	(0.00448)	(0.000226)	(0.000326)
Health - Fair	0.0328***	0.0148	0.0285***	0.0348***	0.00193***	0.00224***
	(0.00516)	(0.0126)	(0.00526)	(0.00787)	(0.000425)	(0.000652)
Cohabit	-0.0492***	0.0445	-0.0353***	-0.0660***	-0.00228***	-0.00398***
	(0.0188)	(0.0331)	(0.0113)	(0.0129)	(0.000749)	(0.000966)
Insurance	-0.00144	0.0177**	-0.00126	-0.00802**	-8.57e-05	-0.000595**
	(0.00321)	(0.00696)	(0.00282)	(0.00386)	(0.000191)	(0.000260)
ln(Wage)	0.00240***	0.00419***	0.00211***	0.000821***	0.000143***	2.87e-05***
	(0.000326)	(0.000572)	(0.000286)	(0.000113)	(2.32e-05)	(4.88e-06)
ln(CPI Drinks)	-0.118*	0.0309	-0.103*	0.00606	-0.00700*	0.000212
	(0.0701)	(0.126)	(0.0615)	(0.0247)	(0.00421)	(0.000862)
ln(Marijuana)	0.0476***	0.0430	0.0418***	0.00842	0.00284***	0.000294
	(0.0174)	(0.0311)	(0.0153)	(0.00608)	(0.00107)	(0.000215)
ln(CPI Cigarette)	0.00118	-0.0368	0.00103	-0.00721	7.00e-05	-0.000252
	(0.0210)	(0.0377)	(0.0184)	(0.00739)	(0.00125)	(0.000259)
Observations	27,583	27,583	27,583	27,583	27,583	27,583

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

CHAPTER 8: CONCLUSIONS AND DISCUSSIONS

8.1. Estimation Concerns for Policy Applications

This chapter discusses the policy implications of the results obtained in chapter 7. The ZIOP model is favored over the OP, the OP-ED, the ZIOPC, and the SSOP models. The ZIOP model accounts for inflated zeros in the endogenous participation problems. This endogenous participation problem is common in models of addictive goods like cigarette and alcoholic beverages. Modeling the impact of first marriage on smoking and drinking outcomes entail endogenous participation problem because the differences between the decision on participation in an activity and the intensity of participation are specified in the ZIOP model as two separate data generating process. Thus, using single equations like the OP model, or the OP-ED model to specify drinking and smoking at the level of zero outcome would not capture the underlying separate data generation process in samples comprising inflated zeros. Misspecification of the ZIOP model will leave out the endogenous participation problem. Thus, wrong smoking and drinking policies can result from these misspecifications.

At least two policy implication can be drawn from the tests of goodness in Chapter 6 and the marginal effects in Chapter 7. First, since individuals must first decide to be a participant or not, a misspecification that leaves out the double hurdle model can understate the benefits of marriage in reducing smoking and drinking. For example, the marginal effects of zero smoking of the "years around first marriage" is 0.0199 under the OP model, whereas this estimate is 0.0203 under the ZIOP model (see Table 7.3). The magnitude of the ZIOP model is bigger than the magnitude of the OP model because the ZIOP model comprises sum of two component parts

(reflecting the two data generating processes) at the level of zero outcome. Similarly, the marginal effect of zero drinking during the "years around first marriage" is 0.0301 under the OP model, whereas it is 0.0333 under the ZIOP model.

Second, one may also be interested in the probability of nonparticipation and the probability of zero consumption conditional on participation. Single equations like the OP, and OP models have no mechanism for decomposing these components of unconditional zero outcomes. Consequently, this decomposition can have consequences for policy implementation. For example, one of the main conclusions in this study is that since $pr(y = 0, w = 1) \approx 0$ for the drinking sample, most drinkers drink regularly albeit lightly or just for socialized drinking. In comparison, $pr(y = 0, w = 1)$ is not close to zero for the smoking sample, and this implies that there are more occasional/infrequent smokers than drinkers. These important sets of result cannot be drawn without expressing the unconditional probability of zero consumption as $pr(y = 0) = pr(w = 0) + pr(y = 0, w = 1)$, a decomposition that cannot be obtained under the OP model.

Simply put, separating the data generating process into participation and consumption decisions creates two categories of policy target groups: abstainers and infrequent users (including recent quitters). Recognizing these two categories of zeros consumption can help policy makers in fashioning different policies for these separate groups. Moreover, the estimates of smoking and drinkers and other special cases of discrete choice model should mimic their true underlying data generating processes.

8.2. Discussions

I examine the effects of first marriage on drinking and smoking outcomes with emphasis on zero consumption in my dissertation. In my models, I emphasized that endogenous participation problem is common in zero consumption of addictive goods like cigarette and alcoholic beverage.

This endogenous problem arises from the differences between intensity of participation and the decision to participate in an activity. The decision to participate in an activity is called the participation decision, while the intensity of participation conditional on the decision to participate is known as the consumption decision.

Consequently, abstention from smoking and drinking is different from infrequent consumptions of these goods. Because of this differences between participation and consumption decisions, the data generating process of outright abstention is different from that of infrequent consumption at zero consumption levels. Thus, modeling these separate sources of "excess" or "inflated" zeros requires a double hurdle model. These excess zeros should not be confused with utility maximizing 'genuine' zeros and corner solution zeros. I applied the zero-inflated ordered probit (ZIOP) model, a double hurdle model, to model the impact of first marriage on smoking and drinking after carrying out a series of model selection tests.

Using the theoretical model of marriage contract, I show that married couples benefit from marriage in terms of reduction in the probabilities of smoking and drinking intensities around and after the year of first marriage. Since marriage is a lifetime commitment with benefits like joint asset ownerships, pooling resources and so on between couples, the marriage market is competitive. As a result, individuals 'clean up' their acts before seeking marriage partners. The cleaning up acts include reduction in smoking and drinking habits. The change in smoking and drinking habits around first marriage can continue well into the marriage union as couples monitor each other. Consequently, marriage can potentially reduce smoking and drinking behaviors.

The main contribution of my dissertation to the literature of risky behaviors and marriage is the estimation of separate probabilities for absolute abstentions and infrequent consumption of tobacco and alcoholic beverage. Key policy implications regarding the probabilities of zero tobacco and alcohol consumption will not be obvious if the smoking and drinking outcomes are

modeled as single equation like the OP model. This is because single equation models like the OP model do not have the mechanism for separating the probability of non-participation from the probability of participation with zero consumption.

Another contribution of this study to the literature is the use of spline function to define different segments of the of the marriage timeline. The slope of the spline function shows smoking and drinking before, around, and after first marriage. Individuals aspiring to get married clean up their acts by reducing their drinking and smoking habits. Because of this, the relative sizes of marriage effects on drinking and smoking habits are not uniform on the baseline regressions. The spline function captures these different impacts of first marriage on smoking and drinking habits at each segment on the function.

I obtain the probabilities of observing zero alcoholic beverage and tobacco consumption in the "years before first marriage", the "years around first marriage", and the "years after first marriage" by estimating the marginal effects of the OP and ZIOP models. Although I show these probabilities for other smoking and drinking outcome intensities, my focus is on the zero outcome. Overall, these marginal effects show that the probabilities of zero consumption of smoking and drinking increases around and after first marriages.

I also find that fewer respondents in the drinking sample are infrequent drinkers. Most of these respondents are either total abstainers or regular/social drinkers. In contrast, there is strong evidence that some respondents in the smoking sample are infrequent smokers. Simply put, there are more infrequent smokers (who belong to the zero-consumption conditional on participation category) than infrequent drinkers in the sample.

Finally, I find that the benefits of first marriage in terms of reduced smoking and drinking is diminishing in consumption levels. That is, the impact of marriage on alcoholic beverage and tobacco consumption wane as one considers higher smoking and drinking intensities. My

empirical results in Tables 7.5 and 7.6 suggest that the benefit of marriage vis-à-vis reduction in tobacco and alcoholic beverage consumption peaked at the ordinal levels of two and one respectively.

`

REFERENCES

Adams, J.D., Chiang, E.P., Jasen, J. 2003. The influence of federal laboratory R&D on industrial research. *Rensselaer Working paper in Economics.* Department of Economics, Rensselaer Polytechnic Institute.

Ali, M.A; Ajilore, O., 2011. Can marriage reduce risky behavior for African-Americans? *Journal of Family Economic Issues*. 32:191-203.

Ariste,I D., Pieroni, L., 2010. A double-hurdle approach to modeling tobacco consumption in Italy. *Applied Economics* 40, 19(2008)

Bachman, J, G., Wadsworth, K.N., O'Malley, Johnston, P.M., L.D., and Schulenburg J.E., 1997. The decline of substance use in young adulthood: changes in social activities, roles, and beliefs. *Mahwah, NJ: Lawrence Erlbaum Associates*.

Becker, S.G., Grossman, M., Murphy, M.K., 1990. An empirical analysis of cigarette addiction. *NBER Working Paper Series* No. 3322

Becker S.G., Murphy M.K, 1988. A Theory of rational addiction. *Journal of Political Economy*. Vol. 96, No 4, pp. 675-700.

Becker, S.G., 1973. A theory of marriage: part 1. *Journal of Political Economy* 81:813-46.

_____, S.G., 1974. A theory of marriage: part 2. *Journal of Political Economy* 82: S11-S26.

Blundell, R.W., Powell, J.L., 2004. Endogeneity in semiparametric binary response models. *Review of Economic Studies* 71, 655-679.

Bratti, M., Miranda, A., 2010. Non-pecuniary returns to higher education: the effect on smoking intensity in the UK. *Health Economics*. 19: 906-920.

Bratti, M., Miranda, A., 2011. Endogenous effects for count data model with endogenous participation or sample selection. *Health Economics*. 20: 1090-1109.

Chaloupka, J.F., 1990. Rational addictive behavior and cigarette smoking. *NBER Working Paper Series* No. 3268

Chaloupka, J.F., Wechsler, H., 1997. Price, tobacco control policies and smoking among young adults. *Journal of Health Economics*, 16, 359-373.

Chaloupka, J.F., Grossman Michael., The demand for cocaine by young adults: a rational addiction approach. *Journal of health Economics* 17, 427-474.

Chesher, A., Smith, R., 1997. Likelihood ratio specification tests. Econometrica 65(3), 627-646

Chester, A., Smolinski, K., 2012. IV models of ordered choice. *Journal of Econometrics*, 166(2012) 33-48.

Cragg, J., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, Vol. 39, No. 5, PP 829-844

Cragg, J.G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* 39, No. 5, 829-844.

De Luca, G., Perotti, V., 2011. Estimation of ordered probit model with sample selection. *Stata Journal*, 11 (2), 213-239.

Duncan, J.G., Wilkerson, B., England, P., 2006. Cleaning up their act: the effects of marriage and cohabitation on licit and illicit drug use, *Demography* Volume 43, Number 4. 2006: 691-710.

Feng, S., 2005. Rationality and self-control: the implication for smoking cessation. *The Journal of Socio-Economics*, 34, 211-222.

Fu, H., Goldman, N., 1996. Incorporating health into models of marriage choice: demographic and sociological perspectives. *Journal of Marriage and the Family* 58: 740-758.

Greene, H.W., 2012. Econometric analysis (Seventh Edition). *Pearson Education Limited.*

Greene, W., 1994. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. Working paper EC-94-10, Stern School of Business, New York University, Stern School of Business, New York University.

Greene, H.W., Hensher, A.D., 2010. Modeling ordered choices: A Primer. Cambridge University Press, Cambridge, UK.

Grossman, M., Chaloupka, F.,1988. The demand for cocaine by young adults: a rational addiction approach. *Journal of Health Economics* 17 (1998) 427-474.

Gurmu, S., Getachew A. D., 2012. Bayesian approach to zero-inflated bivariate ordered probit regression model, with an application to tobacco use. *Journal of Probability and Statistics*. Volume 2012.

Harris, N.M., Zhao, X., 2007. A zero-inflated ordered probit model, with an application to modeling tobacco consumption. *Journal of Econometrics* 121, 1037-1099

Heckman, J., 1979. Sample selection bias as a specification error. *Econometrica* 47, 153-161

Humphreys, B.R., 2013. Dealing with zeros in economic data. University of Alberta, Department of Economics. Retrieved from https://sites.ualberta.ca/~bhumphre/class/zeros_v1.pdf

James, M. A., 1992. A note on computation of the double hurdle model with dependence with an application to tobacco expenditure. Bulletin of Economic Research 44:1.
Jones, M.A., 1989. A double-hurdle model of cigarette consumption. *Journal of Applied Econometrics* 4, No.1, 23-39.

Jones, M. A., Labega M. J., 2003. Individual heterogeneity and censoring in panel data estimates of tobacco expenditure. *Journal of Applied Econometrics*, Vol.18, N0.2, pp. 157-177.

Kenkel, M., Terza, J. 2001. The effect of physician advice on alcoholic consumption: count regression with an endogenous treatment effect. *Journal of Applied Econometrics.* 16(2): 165-184.

Lambert, D., 1992. Zero inflated Poisson regression with an application to defects in manufacturing. Technometrics 34, 1-14.

Lee, M.R., Chassin, L., Mackinnon, D. The effect of marriage on young adult heavy drinking and its mediators: results from two methods of adjusting for selection into marriage. *Psychology of Addictive Behaviors*. Vol.24, No.4, 712-719

Leonard, E. K., Rothbard C.J., 1999. Alcohol and the marriage effect. *Journal of Studies on Alcohol*. Supplement No.13, 1999.

Leonard, K.E., Smith, P.H., Homish, G.G., 2014. Concordant and discordant alcohol, tobacco, and marijuana use as predictors of marital dissolution. *Psychology of Addictive Behaviors*. Vol.28, No.3, 780-789

Lonardo, A. R., Manning, D.W., Giordano, C. P., Longmore, A.M., 2010. Offending, substance use, and cohabitation in young adulthood. *Sociological Forum*, *Vol.25, No.4, December 2010.*

Luca, G.D; Perotti V., 2010. Estimation of ordered response models with sample selection. *CIES TOR Vergata* Research Paper Series, Vol.8. Issue3, No. 168.

Maddala, G.S., 1983. Limited dependent and quantitative variables in econometrics. Cambridge University Press, Cambridge, UK.

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.

Matouschek, N., Rasul, I., 2008. The economics of marriage contract: theories and evidence. *Te journal of law and economics,* Vol. 51, N0.1 (February 2008), pp. 59-110.

Mastekaasa A., 1992. Marriage and psychological well-being: some evidence on selection into marriage. *Journal of Marriage and Family.* Vol.54, No.4, pp. 901-911

Mayer, W.J., Ciu, K., Simona L.T. 2010. An empirical analysis of determinants of binge drinking. *Working paper series, University of Mississippi.*

McKelvy, R., Zavoina, W., 1975. A statistical model for the analysis of ordinal level dependent variables. Journal of Mathematical Sociology 4, 103-120

Miller –Tutzauer., C, K.E. Leonard., and M. Windle., 1991. Marriage and alcohol Use: A longitudinal Study of 'Mating Out'. *Journal of Studies on Alcohol* 52, 434-440.

Mullahey, J., 1997. Heterogeneity, excess zeros and the structure of count data model. *Journal of Applied Econometrics* 12, 337-350.

Murray, J.E., 2000. Marital protection and marital Selection: evidence from a historical perspective sample of American men. *Demography*. 37.4 (2000) 511-521

Paul, M.V., 1974. Over insurance and public provision of insurance: The Roles of Moral Hazard and Adverse Selection. *The Quarterly Journal of Economics*. (1974) 88 (1): 44-62.

Rabe-Hesketh, S., Miranda, A. 2006. Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal and count variables. *The Stata Journal*, Vol.6, Number 3, pp. 285-308.

Roodman, D., 2011. Fitting fully observed recursive mixed-process models with cmp. *The Stata Journal*, Vol. 11, Number 2, pp. 159-206.

Schulenburg, J; O'Malley, P.M., Wadsworth, N.K., Johnston, D.L., 1995. Getting Drunk and Growing Up: Trajectories of Frequent Binge Drinking during the Transition to Young Adulthood, *Survey Research Centre, Institute for Social Research, University of Michigan.*

Weschsler, H., Lee, E. J., Kuo, M., Lee H., 1999. College binge drinking in the 1990s: A continuing problem. Results of the Harvard School of Public Health 1999 College Alcohol Study. *Journal of American College Health*. Vol.48, March 2000.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24-36

Train, K., 2009. *Discrete Choice Methods with Simulations*. Cambridge University Pres: Cambridge, UK.

U.S. Department of Human Services., 2014. The health consequences of smoking- 50 years of progress: A report of the surgeon general. *National Centre for Chronic Disease Prevention and Health Promotion, Office of Smoking and Health*.

Umberson, D., 1992. Gender, marital status and the social control of health behavior. *Social Science and Medicine.* Vol.34, issue 8, pages 907-917

Silva, J.M.C., Tenreyro, S., Windmeijer, F., 2014. Testing non-nested models for non-negative data with many zeros. *Journal of Econometric Methods*, 4(1). Pp. 29-46
Voung, Q., 1989. Likelihood ratio test for model selection and non-nested hypotheses. Econometrica 57, 307-334.

Williams K., Umberson, D., 2004. Marital status marital transition and health: A gendered life course perspective. *Journal of Health and Social Behavior*, Vol. 45, 81-98.

Wilson, P., 2014. The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economic Letters* 127, 51-53.

Wold, S., 1974. Spline function in data analysis. *Technometrics,* Vol. 16 (1), pp. 1-11.

Wooldridge, J.M., 2010. Econometric analysis of cross section and panel data. *MIT Press*, Cambridge, Massachusetts.

World Health Organization., 2015. WHO global report on trends in prevalence of tobacco smoking*: Geneva. World Health Organization.*

APPENDICES

Appendix A

Based on Figure 4.1, the spline specification of the impact of first marriage on smoking is:

$$\begin{aligned} Y_{it}^{0S*} &= {\beta_0}^0 + {\beta_1}^0 t + \underline{\pi^0}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}^0 && \text{if } t < 3 \\ Y_{it}^{0S*} &= {\beta_0}^1 + {\beta_1}^1 t + \underline{\pi^1}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}^1 && \text{if } 3 \le t < 6 \\ Y_{it}^{0S*} &= {\beta_0}^2 + {\beta_1}^2 t + \underline{\pi^2}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}^2 && \text{if } t \ge 6. \end{aligned} \qquad (A.1)$$

Combining all the three equations in equation (1):

$$Y_{it}^{0S} = \alpha_0 + \alpha_1 t + \gamma_2 d_2 + \theta_2 d_2 t + \gamma_3 d_3 + \theta_3 d_3 t + \underline{\pi}\,\underline{X_{it}} + \varepsilon_{it,Y^{0S}}. \qquad (A.2)$$

It should be noted that d_1 is omitted to get a full rank condition in the baseline regression in equation (2). Let $d_1 = 1$ if $t < t_1^*$, $d_2 = 1$ if $t \ge t_2^*$, and $d_3 = 1$ if $t_1^* < t \le t_2^*$ where $t_1^* = 3$ and $t_2^* = 6$. The threshold values $t_1^* = 3$ and $t_2^* = 6$ are the knots. d_1, d_2, and d_3 are marriage dummy variables.

The slopes of the spline function in equation (2) are the parameters of the interactions of first marriage dummy variables and time. These interactions are the impact of time on the probability of smoking in cases involving single or married smokers. Thus, the slopes of the spline function in equation (2) are: α_1, $\alpha_1 + \theta_2$, and $\alpha_1 + \theta_2 + \theta_3$. The 'jumps' or intercepts on Figure 4.1 occur at α_0, $\alpha_0 + \gamma_2$, and $\alpha_0 + \gamma_2 + \gamma_3$.

To make the spline function in equation (2) a continuous function, one needs to join the segments at the knots. For the first knot at $t_1^* = 3$, the knot is joined at:

$$\alpha_0 + \alpha_1 * t_1^* = (\alpha_0 + \gamma_2) + (\alpha_1 + \theta_2) * t_1^*. \qquad (A.3)$$

For the first knot at $t_1^* = 6$, the knot is joined at:

$$(\alpha_0 + \gamma_2) + (\alpha_1 + \theta_2) * t_2^* = (\alpha_0 + \gamma_2 + \gamma_3) + (\alpha_1 + \theta_2 + \theta_3) * t_2^*. \qquad (A.4)$$

From equation (3):

$$\alpha_0 + \alpha_1 * t_1^* = \alpha_0 + \gamma_2 + \alpha_1 * t_1^* + \theta_2 * t_1^*.$$

$$\gamma_2 + \theta_2 * t_1^* = 0.$$

Thus,

$$\gamma_2 = -\theta_2 * t_1^*. \qquad (A.5)$$

Also, from equation (4):

$$\alpha_0 + \gamma_2 + \alpha_1 * t_2^* + \theta_2 * t_2^* = \alpha_0 + \gamma_2 + \gamma_3 + \alpha_1 * t_2^* + \theta_2 * t_2^* + \theta_3 * t_2^*.$$

$$\gamma_3 + \theta_3 * t_2^* = 0.$$

$$\gamma_3 = -\theta_3 * t_2^*. \qquad (A.6)$$

Plugging equation (5), and (6):

$$Y_{it}^{0S} = \alpha_0 + \alpha_1 t + (-\theta_2 * t_1^*) d_2 + \theta_2 d_2 t + (-\theta_3 * t_2^*) d_3 + \theta_3 d_3 t + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0s}}.$$

$$Y_{it}^{0S} = \alpha_0 + \alpha_1 t + \theta_2 d_2 (t - t_1^*) + \theta_3 d_3 (t - t_2^*) + \underline{\pi}\, \underline{X_{it}} + \varepsilon_{it,Y^{0s}}. \qquad (A.7)$$

VITA

NAME:	Lateef A. Subair
PLACE OF BIRTH:	Lagos Island, Lagos State, Nigeria
EDUCATION:	University of Lagos, Lagos, Nigeria 2000-2005 B.Sc.
	University of New Hampshire, New Hampshire 2009-2010 M.A
	University of Mississippi 2010-2013, M.A
	University of Mississippi 2010-2018, Ph.D.
WORK EXPERINCE:	Zenith International Bank, Plc. 2005-2009
	Graduate Assistant, University of New Hampshire 2009-2010
	Graduate Instructor, University of Mississippi 2011-2018