The E®ect of Smoking on Health Using a Sequential Self-selection Model

Kajal Lahiri and Jae G. Song¤

Department of Economics,
State University of New York at Albany,
Albany, NY 12222, USA

Abstract

We estimate a structural model of individual smoking behavior emphasizing the role of individual risk belief on smoking choices. Our model consists of ¯ve equations: two selection equations for initiation and cessation decisions, and three switching outcome regressions for nonsmokers, ex-smokers, and current smokers. The presence of signi¯cant self-selectivity implies that the health e®ects of smoking based on sample proportions do not correctly indicate the true risk of cigarette smoking. Further, our evidence suggests that the self-selection in the cessation decision, but not in the initiation decision, is consistent with economic rationality. We estimate the model by FIML with starting values from heteroskedasticity corrected Heckman-Lee two-step method using newly released Health and Retirement Study(HRS) data.

* Earlier versions of the paper were presented at the 1998 Econometric Society Meetings at Chicago, and the Triangle Health Economics Workshop. We are grateful to Andrew Jones, Terrence Kinal, Hamp Lankford, Maarten Lindeboom, Angel Lopez-Nicolas, Edward Norton, Ilde Rizzo, Michael Sattinger, Frank Sloan, Insan Tunali, and Michelle White for many helpful comments and discussions. We thank Debbie Dwyer for ¯rst suggesting to us the possible use of HRS data in the present context.

1. Introduction

Ever since a causal relationship between cigarette smoking and coronary heart disease was reported at Mayo Clinic in 1940, the effect of cigarette smoking on health has been extensively studied by both epidemiologists and social scientists. Now cigarette smoking is considered to be the number one source of preventable morbidity and premature mortality in the United States (see Bartecchi et al. (1994)). Since the 1960s, various public policy efforts have been devoted to reduce the prevalence of cigarette smoking. Still over 25% of the population, or nearly 46.3 million adults in the U.S. and many more around the world smoke on a regular basis (National Center for Health Statistics (1996)).

There have been two approaches to reduce the prevalence of cigarette smoking: discouraging nonsmokers from initiating, and encouraging smokers to quit. Even though smoking prevalence has steadily decreased since 1960s, this decrease has not been due to fewer people initiating but due to more people quitting (National Center for Health Statistics (1996)). Hence a comprehensive model of smoking participation that considers both initiation and cessation behavior and their health effects is essential for developing an effective public policy approach to reduce the prevalence of smoking, and in estimating the proper human costs of cigarette smoking.

Unlike the consumption of a normal good, cigarette consumption increases not only immediate satisfaction for smokers but also the probability of adverse health effects in the future, neither of which are directly observable. Hence the subjective judgement on the costs and benefits of cigarette smoking plays a crucial role in smoking participation decisions. This subjective judgement largely depends on the

assessment of probability of the occurrence of side e®ects and the time preference between the immediate bene¯ts and the future side e®ects of smoking.

An important factor a®ecting the initial smoking decision is an individual's prior belief on risks and bene¯ts from smoking which could depend on his demographic and other socioeconomic characteristics. Those who decide to smoke gather additional information through their experience of smoking, and update their prior beliefs. Based on the updated belief a decision of continuation or cessation is made. Therefore individual risk assessment of cigarette smoking takes a critical role in each phase of the smoking cycle. Viscusi (1992) studied individual risk perception and smoking decision by questioning whether smokers are risk cognizants, and whether the risk perception is re°ected in smoker's behavior. Viscusi (1991) ¯nds that the risk perception by the young is quite high, but has no signi¯cant in°uence on their initiation behavior.

Recently, the participation behavior of smokers has been empirically examined by Jones (1994,1995) and Hsieh et al. (1996). In these studies, the role of health condition, health knowledge, social interaction, and other demographic characteristics are explored. There is another group of studies which examines the various human costs of smoking, such as health conditions, medical expenditures, and other economic consequences. Miller et al. (1994) estimate medical care expenditures attributable to cigarette smoking. Mattson et al. (1987) calculate the long-term risk of death contributed by individual smoking status for various age groups. The later group of studies, however, treat smoking status as exogenous, disregarding the dynamic interaction between individuals' health conditions and their smoking behavior.

In this paper, we study the participation behavior of smokers, both initiation and cessation, and integrate them into a model of health consequences of smoking. We examine the possibility of the presence of any unmeasured heterogeneity bias in the probability distributions of smoking-related diseases for different smoking groups, and study whether the observed proportions of smoking-related diseases in the sample correctly represent the risk factor associated with a particular smoking behavior. The second issue we examine is whether the individual smoking choices are made in a way that is consistent with economic optimality. The rational addiction approach of cigarette smoking behavior considers smoking choice as the outcome of individual utility maximization under uncertainty (see Becker and Murphy (1988), and Orphanides and Zervos (1995)). Even though there are a number of empirical studies on the rational addiction model, they tend to focus on the role of price on the demand for addictive goods without considering the role of implicit health cost.

The paper is organized as follows. In sections 2 and 3, we develop a sequential selection model of smoking behavior based on a typical smoking cycle. Sections 4 and 5 describe our data and the econometric strategy. Section 6 presents and discusses empirical findings and section 7 concludes the paper.

2. The Econometric Model

The smoking participation decisions, both initiation and cessation, are modeled as outcomes of utility maximization under uncertain occurrence of smoking-related diseases (SRDs) using a random utility model. Our model is based on the following fundamental premise: the baseline (autonomous) and induced risk factors of

cigarette smoking are not always equal for all individuals, and that each individual possesses a subjective prior belief concerning the probability of occurrence of SRDs associated with each smoking choice, and this belief is updated using the information gained through the smoking experience and the realization of changes in his health condition.

Consider a simple two period model with a time separable indirect utility function. A rational individual lives two periods indexed by $t = 1; 2$. At the beginning of each period, individual i faces a decision to make a choice between two alternatives, smoke or not smoke indexed by $j = 1; 2$ based on his own subjective judgement on the costs and the bene¯ts of the alternatives. The length of each period varies across individuals. Let there be two discrete states of health condition, good and bad, indexed by $l = 1; 2$. The bad health state indicates the presence of any undesirable heath condition, caused by factors including smoking. Individual i possesses a prior belief on the probability of the occurrence of each health state for a given smoking status. This probability is optimally updated each period. Let P_{jl} be the individual's subjective probability for l^{th}health state when the smoking status is j. There are four possible states (S_{jl}) of the world an individual could be falling into. Further let U^t_{ijl} be an unobservable indirect utility of individual i choosing alternative j at period t when his/her health condition is l: Then the expected utility of each choice for each period t can be expressed as the sum of the utilities in each health state weighted by its probability :

$$EU^t_{ij} = \sum_{l=1}^{2} P^t_{ijl} U^t_{ijl} \tag{2.1}$$

At the beginning of each period, an individual chooses alternative j over j' if

and only if $E_t U_{ij} > E_t U_{ij'}$. The discounted sum of expected utilities for the two periods realized at the beginning of period ¿ is :

$$E_¿ U_{ij} = \sum_{t=¿}^{¿+1} \mu_i^{t-¿} \sum_{l=1}^{2} P_{ijl}^t U_{ijl}^t \qquad (2.2)$$

where μ_i is an individual-specific time preference parameter.

The initiation decision rule at the beginning of period 1 is governed by the sign of $I_{1i}^* = E_1 U_{i1} - E_1 U_{i2}$. Once an individual starts to smoke then another subsequent decision rule can be defined in a similar way. The cessation decision rule for the second period is governed by the sign of $I_{2i}^* = E_2 U_{i1} - E_2 U_{i2}$. Thus, the selection criteria are:

$$\begin{aligned} \Pr(\text{choose to start smoking}) &= \Pr(I_{1i}^* > 0) \\ \Pr(\text{choose not to start}) &= \Pr(I_{1i}^* \leq 0) \\ \Pr(\text{choose to continue}) &= \Pr(I_{2i}^* > 0; I_{1i}^* > 0) \\ \Pr(\text{choose to quit}) &= \Pr(I_{2i}^* \leq 0; I_{1i}^* > 0) \end{aligned}$$

I_{1i}^* and I_{2i}^* can be interpreted as present values of 'net' utilities at the time of initiation and cessation. They are not observable; we only observe their binary outcomes, I_{1i} and I_{2i}. There are a total three mutually exclusive outcomes of the selection process:

$$\begin{aligned} \text{Group I(nonsmoker, } I_{00}) &: I_{1i} = 0 \\ \text{Group II(ex-smoker, } I_{10}) &: I_{1i} = 1 \text{ and } I_{2i} = 0 \\ \text{Group III(current smoker, } I_{11}) &: I_{1i} = 1 \text{ and } I_{2i} = 1 \end{aligned}$$

The two selection equations are parameterized as:

$$\begin{aligned} I_{1i}^{\alpha} &= Z_{1i}\gamma_1 + W_{1ij}\theta_1 + \varepsilon_{1i} \\ I_{2i}^{\alpha} &= Z_{2i}\gamma_2 + W_{2ij}\theta_2 + H_{2i}\delta_2 + \varepsilon_{2i} \end{aligned} \tag{2.3}$$

where $(Z_{1i}; Z_{2i})$ are individual characteristics and $(W_{1ij}; W_{2ij})$ are the characteristics of the alternatives specific to the individual. The updated probability assessment at the beginning of the second period depends on the realization of change in health conditions (H_{2i}) which may or may not be related to the smoking status in period 1, individual characteristics (Z_i), and characteristics of the alternatives specific to the individual (W_{ij}).

Finally, we specify an equation for the appropriate response variable for measuring the effect of the initiation and the cessation decisions. Since the smoking participation decisions heavily influence the probability of SRDs such as lung diseases and various types of cancer, presence of any SRDs is the most appropriate and direct outcome variable for our purpose. Further, we will specify one equation for each smoking group in order to capture the full interactions among them. The equations are:

$$\begin{aligned} Y_{ni}^{\alpha} &= X_{ni}\beta_n + \varepsilon_{ni} : \text{ nonsmokers' disease equation} \\ Y_{xi}^{\alpha} &= X_{xi}\beta_x + \varepsilon_{xi} : \text{ ex-smokers' disease equation} \\ Y_{ci}^{\alpha} &= X_{ci}\beta_c + \varepsilon_{ci} : \text{ current smokers' disease equation} \end{aligned} \tag{2.4}$$

In (2.3)-(2.4) Z_{1i}; and W_{1ij} are $N \times K_1$ and $N \times K_2$ vectors of explanatory variables; γ_1; and $\theta_1$are $K_1 \times 1$;and $K_2 \times 1$ vectors of unknown coefficients; $Z_{2i}; W_{2ij}$; and $H_{2i;}$ are $N_1 \times K_3; N_1 \times K_4$;and $N_1 \times K_5$ vectors of explanatory variables; and

$\acute{}_2; \pm_2$;and $^3_2$ are $K_3 \pounds 1$; $K_4 \pounds 1$;and $K_5 \pounds 1$ vectors of unknown coe±cients; $X_{ni}; X_{xi}$; and X_{ci} are $N_2 \pounds K_6; N_3 \pounds K_6$;and $N_4 \pounds K_6$ vectors of explanatory variables; $\bar{}_n; \bar{}_x$;and $\bar{}_c$ are $K_6 \pounds 1$ vector of unknown coe±cients; I^{α}_{1i}; I^{α}_{2i}; Y^{α}_{ni}; Y^{α}_{xi};and Y^{α}_{ci} are $N \pounds 1$; $N_1 \pounds 1$; $N_2 \pounds 1$; $N_3 \pounds 1$; and $N_4 \pounds 1$ unobservable latent indices; $N = N_2 + N_3 + N_4$;and $N_1 = N_3 + N_4$. We observe binary outcomes Y_{ni}; Y_{xi}; Y_{ci}; I_1;and I_2:

This completes our econometric model as a set of switching regressions with sequential self-selection rules. In the remainder of this paper subscript i is omitted to avoid notational complexity. Also new variables $C_1$and C_2 will represent the independent variables in the ¯rst and the second selection equations in (2.3) respectively with coe±cients $°_1$and $°_2$.

$I^{\alpha}_1 = C_1°_1 + ²_1$: initiation selection equation

$I^{\alpha}_2 = C_2°_2 + ²_2$: cessation selection equation (observed, i® $I_1 = 1$)

$Y^{\alpha}_n = X\bar{}_n + ²_n$: nonsmoker's disease equation (observed, i® $I_1 = 0$)

$Y^{\alpha}_x = X\bar{}_x + ²_x$: ex-smoker's disease equation (observed, i® $I_1 = 1; I_2 = 0$)

$Y^{\alpha}_c = X\bar{}_c + ²_c$: current smoker's disease equation (observed, i® $I_1 = 1; I_2 = 1$)

The two sequential self-selection rules sort people into observed classes according to the expected present value of indirect utility. Hence the presence of SRDs actually observed in each group are not random outcomes in the population, but instead are self-censored nonrandom samples. The initiation decision equation is de¯ned over the entire population, but the cessation decision equation is de¯ned only over the subset of observations for those who have started to smoke.

3. Data and Empirical Speci¯cations

We use data from Health and Retirement Study (HRS) Wave I which was released in May 1995. The HRS is a national longitudinal study on health, retirement, and economic status focusing on individuals born between 1931 and 1941. A total of 12,652 individuals were interviewed during 1993, among these 2,373 are single respondents and 5,234 are paired (married or partnered) respondents. Their ages vary from 23 to 85 as of 1993. Mean age in the total sample is 55.6 and standard deviation is 5.66. Out of 12,652 individuals, 4,626 are nonsmokers, 4,588 are ex-smokers, and 3,438 are current smokers.

The main advantage of this data set for our purpose is that the sample mainly consists of individuals in their 50s. The role of learning and regret throughout one's smoking cycle and their e®ects on health can be more fully observed in this data set because most smokers initiate smoking when they are relatively young. The HRS also provides complete classi¯cation of individual smoking status as nonsmoker, ex-smoker, and current smoker. Further, HRS has extraordinary information on both current health conditions, health history, and various socio-economic factors. As a result, we are able to investigate the long-term health e®ects 30-40 years of smoking in a comprehensive manner.

To take full advantage of such characteristics of the data set for our purpose, we consider only individuals in the age group 52 and over. We also drop those individuals who had quit smoking after they were diagnosed for various types of cancers and other smoking-related diseases to control for obvious endogeneity. As a result of these, our ¯nal sample for empirical analysis consists of 9,109 individuals among them 3,287, 3,368, and 2,454 are nonsmokers, ex-smokers, and

current smokers respectively[1].

The HRS data provides information that separate never smokers from ever smokers, and ever smokers are further subdivided into ex-smokers and current smokers. Technically, nonsmokers in HRS are de¯ned as those individuals who smoked less than 100 cigarettes in his/her entire life time. Even though we observe complete outcomes of the two participation decisions, estimation of these equations based on a single cross sectional data requires careful thoughts. Fortunately the data contains large amount of information on current as well as historical health, socio-economic and demographic factors. Our variables are categorized as pure demographic, social status, economic status, family life, current health condition and history, risk taking behavior, smoking-related, and employment related variables. Some of these variables represent time invariant individual characteristics, such as sex, race, place of birth, parents' education, etc. We also use some other variables which are assumed to capture individual characteristics when they were young such as religion, tidiness, smoking status of the spouse, occupation, etc. This is particularly useful for our empirical model speci¯cation because our model contains two decisions which were possibly made many years ago.

We select the explanatory variables for each equation based on previous empirical speci¯cations, theory, and the data availability. From an economic standpoint, the individual's risk belief is assumed to have an important in°uence on the smok-

[1]We drop total of 3,543 individual from our ¯nal sample: 2,407 individuals whose age is less than 52, 269 individuals who quit smoking after they were diagnosed for various diseases, 96 individuals whose household level information is not available, and the rest who have missing information in any one of our variables. Arguably, 269 smokers who quit after being diagnosed having SRD should be treated as current smokers for calculating the risk factor for this group. This will result in a slightly higher risk factor for current smokers than the one reported later in the paper.

ing behavior. Barsky et al. (1997) show that individual risk tolerance measured by HRS is positively related to risk-taking behavior in smoking and drinking. Their estimated preference parameters are related to the behavior of individuals, and their risk tolerance estimates make correct prediction of smoking behavior at least qualitatively. Viscusi (1990,1991) found some evidence that an individual's smoking decision responds to his/her risk perception. Di®erences in smoking behavior among di®erent demographic groups are reported in various studies. The Surgeon Generals' report of 1985 noted di®erences in smoking behavior, both initiation and cessation, between white-collar and blue-collar workers. Breslau and Peterson (1996a, 1996b) reported that smoking cessation varies by sex, race, education, and number of cigarettes smoked daily. They also found that smoking and drinking habits are correlated. The prevalence of alcohol abuse or dependence was signi¯cantly higher in smokers than nonsmokers. For the three SRD equations, we include occupation, smoking, nutrition, demographic, socio-economic, and general health related variables. To control for individual heterogeneity further, we include variables related to occupation, occupational hazards, fundamental health condition indicators, socio-economic status, life-styles, and insurance status. Our key dependent variable is the presence of one of the SRDs which include various lung diseases and types of cancers that are directly related to smoking such as cancer of the abdomen, mouth, bladder, neck, nose, pancreas, brain, bronchia, cervix, esophagus, stomach, throat, tongue, kidney, liver, and lung. Selection of such cancers is based on reports of the Surgeon General in 1982, 1983, and 1984 on the health consequences of smoking, and various medical and public health literature such as Yuan et el. (1996), Bartecchi et el. (1994), Mattson et el.

(1987), and Fielding (1985). Thus, our de¯nition of SRD is very broad, giving us a reasonable sample size of people having SRDs. Detailed de¯nitions of variables are shown in Appendix 2 and descriptive statistics by smoking status for some key variables are summarized in table 1.

4. Empirical Strategy

Our empirical strategy is to estimate the model ¯rst by two-step probit method using Heckman-Lee two-step method, and use the two-step estimates as starting values for full information maximum likelihood (FIML) estimation. The conditional expectations of the dependent variables using the properties of truncated normal density functions are :

$$
\begin{aligned}
E(I_1) &= G_1 \\
E(I_2 \;|\; I_1 = 1) &= G_2 + E(\epsilon_2 \;|\; -G_1 < \epsilon_1) \\
&= G_2 + \rho_{12}\frac{\acute{A}_1(G_1)}{\copyright_1(G_1)} \\
E(Y_n \;|\; I_1 = 0) &= X_n\bar{}_n + E(\epsilon_n \;|\; -G_1 \leq \epsilon_1) \\
&= X_n\bar{}_n + \rho_{1n}\frac{-\acute{A}_1(G_1)}{1 - \copyright_1(G_1)} \\
E(Y_x \;|\; I_1 = 1; I_2 = 0) &= X_x\bar{}_x + E(\epsilon_x \;|\; -G_1 < \epsilon_1; -G_2 \leq \epsilon_2) \\
&= X_x\bar{}_x + \rho_{1x}\frac{\acute{A}_1(G_1)\copyright_1(-G_2^{\ddagger})}{\copyright_2(G_1; -G_2; -\rho_{12})} + \rho_{2x}\frac{-\acute{A}_1(G_2)\copyright_1(G_1^{\ddagger})}{\copyright_2(G_1; -G_2; -\rho_{12})} \\
E(Y_c \;|\; I_1 = 1; I_2 = 1) &= X_c\bar{}_c + E(\epsilon_c \;|\; -G_1 < \epsilon_1; -G_2 < \epsilon_2) \\
&= X_c\bar{}_c + \rho_{1c}\frac{\acute{A}_1(G_1)\copyright_1(G_2^{\ddagger})}{\copyright_2(G_1; G_2; \rho_{12})} + \rho_{2c}\frac{\acute{A}_1(G_2)\copyright_1(G_1^{\ddagger})}{\copyright_2(G_1; G_2; \rho_{12})}
\end{aligned}
$$

where $G_1 = C_1 \circ \sigma_1; G_2 = C_2 \circ \sigma_2; G_1^{\ddagger} = \frac{G_1 - \rho_{12}G_2}{(1-\rho_{12}^2)^{\frac{1}{2}}}$; $G_2^{\ddagger} = \frac{G_2 - \rho_{12}G_1}{(1-\rho_{12}^2)^{\frac{1}{2}}}$; $\copyright_g(\Diamond)$ and $\acute{A}_g(:)$ are the standard g-variate normal distribution and density function respectively.

Thus, we can rewrite the equations with new error terms which have zero conditional means:

$$\begin{aligned} I_1^{\alpha} &= G_1 + {}^{2}{}_1 \\ I_2^{\alpha} &= G_2 + {}^{3}\!/\!_{4}{}_{12}\,{}_{,12} + e_2 \\ Y_n^{\alpha} &= X_n\,{}^{-}{}_{n} + {}^{3}\!/\!_{4}{}_{1n}\,{}_{,1n} + e_n \\ Y_x^{\alpha} &= X_x\,{}^{-}{}_{x} + {}^{3}\!/\!_{4}{}_{1x}\,{}_{,1x} + {}^{3}\!/\!_{4}{}_{2x}\,{}_{,2x} + e_x \\ Y_c^{\alpha} &= X_c\,{}^{-}{}_{c} + {}^{3}\!/\!_{4}{}_{1c}\,{}_{,1c} + {}^{3}\!/\!_{4}{}_{2c}\,{}_{,2c} + e_c \end{aligned} \tag{4.1}$$

where ${}_{,12} = \frac{\acute{A}_1(G_1)}{\copyright_1(G_1)}$; ${}_{,1n} = \frac{\text{¡}\,\acute{A}_1(G_1)}{1\,\text{¡}\,\copyright_1(G_1)}$; ${}_{,1x} = \frac{\acute{A}_1(G_1)\copyright_1(\text{¡}\,G_2^{\alpha})}{\copyright_2(G_1;\,\text{¡}\,G_2;\,\text{¡}\,{}^{3}\!/\!_{4}{}_{12})}$; ${}_{,1c} = \frac{\acute{A}_1(G_1)\copyright_1(G_2^{\alpha})}{\copyright_2(G_1;G_2;{}^{3}\!/\!_{4}{}_{12})}$; ${}_{,2x} = \frac{\text{¡}\,\acute{A}_1(G_2)\copyright_1(G_1^{\alpha})}{\copyright_2(G_1;\,\text{¡}\,G_2;\,\text{¡}\,{}^{3}\!/\!_{4}{}_{12})}$; and ${}_{,2c} = \frac{\acute{A}_1(G_2)\copyright_1(G_1^{\alpha})}{\copyright_2(G_1;G_2;{}^{3}\!/\!_{4}{}_{12})}$. Under the normality assumption, the whole model can be estimated by ¯ve separate probit regressions. Note that the error terms e_2, e_n, e_x, and e_c are heteroskedastic by construction, and if there are overlapping variables in the selection and disease equations, the problem of signi¯cant multicollinearity may result in the structural equations.

4.1. Endogeneity of Some Health Variable

A recent debate between Jones (1994, 1996), and Shmueli (1996) deals with the issue of endogeneity of health variables in the cessation decision equation. The health variables included in our equations are more objective based on physical activity limitations and health history, which are expected to develop independent of smoking.

Blundell and Smith (1986) provide a simple way to test for endogeneity by a two stage approach. The ¯rst stage is to regress the suspicious variables on exogenous variables. The second step is to estimate the original probit equation with

the residuals from stage 1 as additional regressors, and jointly test the hypothesis that coe±cients of the residuals are zeros. Since our suspicious health variables are binary, we generated probit generalized residuals (see Gourieroux et al. (1987) and Vella (1993)). The probit generalized residuals for a model $Y_i^{\alpha} = X_i\bar{}_i + "_i$ are given by:

$$\text{ı}^{\wedge} = (Y_i \text{ ¡ } ©_1(X_i\text{-}^{\wedge}_i))Á_1(X_i\text{-}^{\wedge}_i)((1 \text{ ¡ } ©_1(X_i\text{-}^{\wedge}_i))^{i\,1}©_1(X_i\text{-}^{\wedge}_i)^{i\,1} \qquad (4.2)$$

We tested the endogeneity of EXER (exercises regularly), JOGAMILE (can jog a mile with no di±culty), BLOODPRS (blood pressure), CHOLSTRL (cholesterol), ARTHRTS (arthritis), and DIABTS (diabetes) in the cessation decision and BLOODPRS, CHOLSTRL, ARTHRTS, and DIABTS in the switching disease equations[2]. The $Â^2$ statistics that these additional coe±cients are zero for cessation, non-smokers', ex-smokers' and current smokers' equations are obtained as 6,585 (df. = 6), 3.474 (df. =4), 3.232 (df. =4) and 3.123 (df.= 4) respectively. These are not statistically signi¯cant at the 5% level of signi¯cance. Our speci¯cations originally included a few other health variables but we dropped them because they did not pass the endogeneity test[3].

[2]We use following variables as exogenous regressors for the ¯rst stage probit regression: MALE, FORNBORN, WESTB, RACEW, RACEB, RACEA, SCHLYRS, SCHLYRS2, PAREDU, CATHOLIC, MILIT, MARID1ST, HOUSE, NOMARAGS, RISK, RELIGS, AGE and a few other obvious exogenous variables like the regional dummies.

[3]The variables were: self-reported current health status, change in health condition during last year, presence of various limitations on daily activities, presence of asthma disease, and heart attack.

4.2. Heteroskedasticity

We examine the presence of heteroskedasticity using the following formulation (Greene 1993):

$$Var("_i) = \sigma_i^2 = [\exp(V_i\,!)]^2$$

$$Y_i^{\ast} = X_i^{-}\beta + "_i \tag{4.3}$$

where V_i is $1 \times p$ vector of observations on a subset of variables, and $!$ is a vector of corresponding parameters. Then log of likelihood function is:

$$\ln l = \sum_i Y_i \ln\left(1 - \frac{X_i^{-}\beta}{\exp(V_i\,!)}\right) + (1 - Y_i)\ln\left[1 - \left(1 - \frac{X_i^{-}\beta}{\exp(V_i\,!)}\right)\right] \tag{4.4}$$

Once this likelihood function is maximized, we can easily check for heteroskedasticity by the likelihood ratio test because $! = 0$ implies homoskedasticity. Further we can identify the special structure of heteroskedasticity and are able to correct it by feasible GLS (see Yatchew and Griliches (1985)). The results indicated that significant heteroskedasticity exists in all our equations at the 5% level. As a result, we specified our equations after allowing for heteroskedasticity:

$$\begin{aligned}
I_1^{\ast} &= D_1 + "_1 \\
I_2^{\ast} &= D_2 + \sigma_{12}\gamma_{12}^{\ast} + "_2 \\
Y_n^{\ast} &= D_n + \sigma_{1n}\gamma_{1n}^{\ast} + e_n \\
Y_x^{\ast} &= D_x + \sigma_{1x}\gamma_{1x}^{\ast} + \sigma_{2x}\gamma_{2x}^{\ast} + e_x \\
Y_c^{\ast} &= D_c + \sigma_{1c}\gamma_{1c}^{\ast} + \sigma_{2c}\gamma_{2c}^{\ast} + e_c
\end{aligned} \tag{4.5}$$

where $D_1 = \frac{C_1\alpha_1}{\exp(V_1w_1)}$; $D_2 = \frac{C_2\alpha_2}{\exp(V_2w_2)}$; $D_n = X_n^{-}\beta_n$; $D_x = \frac{X_x^{-}\beta_x}{\exp(V_xw_x)}$; $D_c = \frac{X_c^{-}\beta_c}{\exp(V_cw_c)}$; and V^{0}s are the variables which condition heteroskedasticity for each equation. $\gamma^{\ast}$s are defined in the same way as in (4.1) with G^{0}s replaced by D^{0}s.

4.3. Normality Tests

Pagan and Vella (1989) derived a normality test for the Tobit model with selectivity which can be directly applied to our probit model with selectivity. Since our cessation and nonsmokers' disease equations involve a single selection, we can apply the test developed by Pagan and Vella with minor modi¯cations. However, we want to estimate our structural model by full information maximum likelihood (FIML), which requires evaluation of trivariate CDFs. As a result, we use the normality test based on Edgeworth expansion of CDF suggested by Lee (1984), generalized to the trivariate case by Lahiri and Song (1999). Here we test for the bivariate normality of $(\epsilon_1; \epsilon_n)$ and the trivariate normality of $(\epsilon_1; \epsilon_2; \epsilon_c)$ and $(\epsilon_1; \epsilon_2; \epsilon_x)$. The $\hat{A}^2$ statistics were calculated as 12.97 (df = 9), 27.42 (df.= 25), 31.64 (df.= 25) respectively. Since these values are less than the critical $\hat{A}^2$ values at the 5% level of signi¯cance, we did not reject the multivariate normality assumption in our context. We should point out that without the heteroskedasticity correction, the normality assumption would have been resoundingly rejected in our sample.

4.4. Full Information Maximum Likelihood Estimation

The ine±ciency of the two-step method relative to maximum likelihood has been critically examined by Nelson (1984), who suggested the more di±cult MLE. The poor performance of Heckman-Lee two-step method is also reported in Nawata and Nagase (1996) and Stolzenberg and Relles (1997). Their results, based on Monte Carlo and empirical examples, suggest that the two-step method may not be a dependable estimator when there is strong multicollinearity between indepen-

dent variables(X) and selectivity correction terms(λ's): If there are no overlapping variables in the selection and the regression equations, then the multicollinearity may not be very high. Since we have overlapping variables, especially in the second selection equation, we can not rule out the possibility of significant multicollinearity. As a result, it is advisable to estimate the model by FIML for correct inferences. See Leung and Yu (1996) for further discussion on this issue.

The log of likelihood function for the model (4.5) is:

$$\begin{aligned}
&\ln L(\beta_1, \lambda_1, \beta_2, \lambda_2, \beta_n, \beta_x, \lambda_x, \beta_c, \lambda_c, \rho_{12}, \rho_{1n}, \rho_{1x}, \rho_{1c}, \rho_{2x}, \rho_{2c}) \\
= &\sum_{i=1}^{N} f(1-I_1)Y \ln \Phi_2(-D_1\beta_1 - D_n\beta_n - \rho_{1n}\lambda_1) + (1-I_1)(1-Y)\ln \Phi_2(-D_1\beta_1 - D_n\beta_n, \rho_{1n}) \\
&+ I_1(1-I_2)Y \ln \Phi_3(D_1\beta_1 - D_2\beta_2 - D_x\beta_x - \rho_{12}\lambda_2 - \rho_{2x}\lambda_2, \rho_{1x}\lambda_1) \\
&+ I_1(1-I_2)(1-Y)\ln \Phi_3(D_1\beta_1 - D_2\beta_2 - D_x\beta_x - \rho_{12}\lambda_2, \rho_{2x}\lambda_2 - \rho_{1x}\lambda_1) \\
&+ I_1 I_2 Y \ln \Phi_3(D_1\beta_1 - D_2\beta_2 - D_c\beta_c - \rho_{12}\lambda_2, \rho_{2c}\lambda_2, \rho_{1c}\lambda_1) \\
&+ I_1 I_2(1-Y)\ln \Phi_3(D_1\beta_1 - D_2\beta_2 - D_c\beta_c - \rho_{12}\lambda_2, \rho_{2c}\lambda_2, \rho_{1c}\lambda_1) \qquad (4.6)
\end{aligned}$$

where $\Phi_2(\cdot,\cdot,\cdot)$; and $\Phi_3(\cdot,\cdot,\cdot,\cdot,\cdot)$ are the standard bivariate and trivariate densities respectively. The parameters are $\beta_1, \beta_2, \beta_3, \beta_n, \beta_x, \beta_c$ and 5×5 covariance matrix of disturbances, Σ:

$$\Sigma = \begin{bmatrix} 1 & \rho_{12} & \rho_{1n} & \rho_{1x} & \rho_{1c} \\ \rho_{12} & 1 & \rho_{2n} & \rho_{2x} & \rho_{2c} \\ \rho_{1n} & \rho_{2n} & 1 & \rho_{nx} & \rho_{nc} \\ \rho_{1x} & \rho_{2x} & \rho_{nx} & 1 & \rho_{xc} \\ \rho_{1c} & \rho_{2c} & \rho_{nc} & \rho_{xc} & 1 \end{bmatrix}$$

Note that $\rho_{2n}, \rho_{nx}, \rho_{nc}$; and ρ_{xc} do not explicitly appear in the likelihood function. It is well known that ρ_{nx}, ρ_{nc};and ρ_{xc} are not identified because the likelihood function does not depend on these parameters (for further discussion on this issue see

Koop and Poirier (1997)). Another parameter, σ_{2n}; is also not identified because the second selection equation is irrelevant for the non-smoker group, and hence it is not in the likelihood function. This is because our two selection equations classify the sample not into four categories but into three categories. Hence an additional restriction on the variance covariance matrix is required for identification. The rest of the parameters | $(K_1 + K_2) + 6 + (K_3 + K_4 + K_5) + 7 + 3K_6$ elements of regression coefficients $\beta_1; w_1; \beta_2; w_2; \gamma_n; \gamma_x; \lambda_x; \gamma_c$ and λ_c; plus six elements of the covariance matrix $\sigma_{12}; \sigma_{1n}; \sigma_{1x}; \sigma_{1c}; \sigma_{2x};$ and σ_{2c} | are identifiable. It is well known that the likelihood function of multivariate probit model is not globally concave unlike that of univariate probit model. The complexity of likelihood function makes the full information maximum likelihood estimation difficult, and there is no ready-made guarantee that one has reached the global maximum.

We estimate the model by FIML with starting values from the Heckman-Lee two-step method with pre-tested forms of heteroskedasticity using GAUSS (version 3.2.4). FIML has seldom been used to estimate switching regression models with double selection because of computational difficulty. One useful tip for the estimation to save significant amount of computing time is to arrange the data by each of the six over-overlapping groups and maximize a form of the log of likelihood function given in (4.7) rather than (4.6):

$$
\begin{aligned}
&\ln L(\beta_1; \lambda_1; \beta_2; \lambda_2; \gamma_n; \gamma_x; \lambda_x; \gamma_c; \lambda_c; \sigma_{12}; \sigma_{1n}; \sigma_{1x}; \sigma_{1c}; \sigma_{2x}; \sigma_{2c}) \\
= &\sum_{I_1=0;Y=1} \ln \Phi_2(-D_1; D_n; -\sigma_{1n}) + \sum_{I_1=0;Y=0} \ln \Phi_2(-D_1; -D_n; \sigma_{1n}) \\
&+ \sum_{I_1=1;I_2=0;Y=1} \ln \Phi_3(D_1; -D_2; D_x; -\sigma_{12}; -\sigma_{2x;}\sigma_{1x})
\end{aligned}
$$

$$+ \sum_{I_1=1;I_2=0;Y=0} \ln \Phi_3(D_1; -D_2; -D_x; -\rho_{12}; \rho_{2x}; -\rho_{1x})$$
$$+ \sum_{I_1=1;I_2=1;Y=1} \ln \Phi_3(D_1; D_2; D_c; \rho_{12}; \rho_{2c}; \rho_{1c})$$
$$+ \sum_{I_1=1;I_2=1;Y=0} \ln \Phi_3(D_1; D_2; D_c; \rho_{12}; \rho_{2c}; \rho_{1c})\} \quad (4.7)$$

If one wishes to maximize the likelihood function as written in (4.6), GAUSS will evaluate two bivariate normal CDFs and four trivariate normal CDFs for each observation. In our estimation, we have saved more than 3/4th.of computing time by maximizing (4.7). We also found that a 'good' starting value is critical to achieve smooth convergence. Further we normalized all continuous variables by their means to prevent any possible interruption during the maximization of the likelihood function. We use Berndt-Hall-Hall-Hausman (BHHH) algorithm for the FIML estimation and it took 29 iterations (with running time approximately 673 minutes in a 400 Htz PC) to converge with a pre-set tolerance level of 0.00001[4]. The variance - covariance matrix of the estimated structural parameters is obtained from the ¯nal information matrix on convergence. Interestingly, the estimates from two stage method and FIML were very close except for the selectivity correction terms[5].

[4]We found that BHHH algorithm tends to converge much faster than other Quasi-Newton methods, such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) or Davidon-Fletch-Powell (DFP). In our experiments, BHHH took nearly 30 iterations whereas BFGS and DFP took more than 150 iterations to converge at the tolerance level 0.00001 with the starting values from the Heckman-Lee two-stage method.

[5]In order to make sure that our estimates maximize the likelihood function globally, we experimented with di®erent starting values and also the method of simulated annealing (see Go®e et al. (1994)). Since this estimation can take very long time, we use FIML estimates as starting values for the simulated annealing method.

5. Empirical Estimates

5.1. Selection Equations

Table 2 presents heteroskedasticity corrected two-step and FIML estimates of the first selection equation together with marginal effects and odd ratios associated with different explanatory variables. To check for goodness-of-fit we compute various measures of pseudo R^2. Among those McKelvey and Zavonia's R^2 is 0.23, and the correct prediction rate is about 70 percent.

The initiation decision varies by different demographic characteristics, males tends to initiate more often than females. Participation in regular religious services (RELIGS), stable marriage life (MARID1ST), and nonsmoking spouse (SPOUSNSM) have high negative correlation with initiation. Individual drinking behavior (EVDRINK, EXDRINK) has a significant positive effect. We also find that the propensity to initiate varies across different ethnic groups and education levels. SCHLYRS (school years) has a significant nonlinear effect on the initiation decision. The variables RISKAVER, MYOPIC, and INVEST represent individuals' attitude toward risk and play a major role in initiation. The variable RISKAVER represents individual risk aversion and MYOPIC represents an aspect of individual's time preference. For example, the individual with MYOPIC= 1 may be considered as a short sighted individual who tends to prefer the immediate benefits of smoking over costs of health deterioration in the future. These risk variables, which we can take to be largely time-invariant, turned out to have substantive impact on the initiation decision; odd ratios and marginal effects of RISKAVER, MYOPIC, and INVEST are (0.92, 1.07, 0.86) and (-0.029, 0.022, -0.05) respectively. This evidence indicates that risk averse individuals tend not

to initiate, and an individual with higher time preference for the immediate tend to initiate more often. Other socio-economic status variables have signi¯cant explanatory powers to explain individual's initiation decisions. The direction of contributions from those variables are consistent with previous studies.

Table 3 presents estimates from heteroskedasticity corrected two-step estimator of the second selection equation (the continuation decision) by FGLS and FIML procedures together with the marginal e®ects and odd ratios. McKelvey-Zavonia's R^2 is 0.31 and the correct prediction rate is again about 70 percent. In terms of the R^2 measure, the ¯t of the second selection equation is slightly better than that of the ¯rst selection equation. As one would expect, spouse's cessation decision (SPOSEXSM) has very strong e®ect on individual's cessation decision. Spouse's health status, and the number of children in the household were not signi¯cant. An interesting ¯nding is that the presence of current good health conditions (EXER, JOGAMILE) and also bad health conditions (BLOODPRS, CHOLSTRL, DIABTS) have strong positive e®ects on the propensity for cessation. When people realize that they have developed some bad health conditions which might be aggravated by smoking, they tend to quit (see Shmueli (1996), and Jones (1996)). Current smokers who still enjoy good health sometimes quit in order to maintain the good health. This evidence simply implies that there are two groups of ex-smokers in our sample - one group quits in order to restore better health (curative way), and the other group quits in order to maintain good health (preventive way).

In addition to these ¯ndings, risk variables in the cessation decision equation also have very interesting implications. RISKAVER, MYOPIC, IRA, LIFEINS,

and HELTHINS are the variables that can capture individual risk behavior. The variable RISKAVER represents individual risk aversion and MYOPIC represents an aspect of individual's time preference. Our results show that those who have higher time preference for the current tend to continue to smoke. The variables LIFEINS and HELTHINS represent individual attitude towards health risk and these variable also have a meaningful interpretation. Also, individuals who have individual retirement accounts (IRA) can be considered as more ¯nancially well planned and the variable IRA has positive contribution to the cessation decision. Schooling (SCHLYRS, SCHLYRS2) and Body Mass Index (BMI) variables have signi¯cant nonlinear e®ect on the cessation decision. Another interesting variable is the cigarette addiction variable (ADDICTION) which is designed to capture the approximate strength of smokers' cigarette addiction. As one would expect, our result indicates that the stronger the addiction, harder it is to quit. Also a stable marriage life (MARRIED, NOMARAGS) and drinking habit (ALCHOLIC) have signi¯cant e®ects on the cessation decision. We also observe variations in cessation propensity by di®erent demographic, occupational, and economic classes, and they are consistent with prior ¯ndings.

5.2. Switching Disease Equations

Tables 4, 5, and 6 present estimates of the disease equations by two-step and FIML methods for the three smoking status groups. We specify our switching disease equations using a number of demographic, occupational, economic, and fundamental health condition variables. One of the interesting variables is HAZ-WORK, which represents occupational exposure to hazards. In some sense, HAZ-WORK also represents individual risk taking behavior as well. The occupational

exposure to hazards has signi¯cant e®ect on the probability of the occurrence of the smoking-related diseases for all three smoking groups. There could be direct and indirect e®ects of this variable. The direct e®ect is the contribution from the exposure; the indirect e®ect comes from the fact that individuals with persistent occupational exposure are often risk-lovers, and such attitude can have a positive e®ect on the probability of smoking-related diseases. Another interesting variable is FEDINS which indicates health insurance coverage by various federal health insurance programs, such as Medicaid and VA. This variable mostly captures people having low socio-economic status. It has a positive e®ect on the probability of having SRDs for all three smoking groups. The marginal e®ect of FEDINS for nonsmokers, ex-smokers, and current smokers are 0.03, 0.03, and 0.08 respectively. Thus, these estimates suggest that smoking by individuals has a considerable social cost as well. Unstable marriage life (NOMARAGS) also has a signi¯cant impact on the presence of the diseases for all three groups of smokers. We also ¯nd demographic variations in the incidence of SRDs.

One of the most signi¯cant ¯ndings of this study is the presence of signi¯cant selectivity coe±cients in previous smokers' and current smokers' disease equations. The selectivity coe±cients in FIML estimation in some cases turned out to be substantially di®erent from those in two-step estimations underscoring the importance of FIML approach in these kinds of models. The selectivity coe±cient in the nonsmokers' equation turned out to be insigni¯cant. The statistically signi¯cant selectivity coe±cients in the last two disease equations imply the endogenous nature of switching in our structural model. We should emphasize that we took utmost care to fully specify our ¯ve equations such that the signi¯cance of the se-

lectivity terms is not due to the omission of observable explanatory variables. For instance, interaction of a number of statistically signi¯cant explanatory variables were not found to be important in the speci¯cations.

5.3. Further Implications of Our Findings

The model estimated in this paper provides a way to identify the true risk of smoking by correcting the self-selection problem in the observed proportions of SRDs in di®erent smoking groups. Since our estimates indicate that selectivity biases are present in both ex-smokers' and a current smokers' disease equations, the risk factor for smokers may be di®erent from that of nonsmokers, even if smokers had never smoked. The observed proportions based on our sample indicate that the probability of getting SRDs for ex-smokers' and current smokers' are 9.41% and 15.97%, respectively. These observed relative frequencies of SRDs in the sample are biased estimates of the true risk factors for the ex-smoker and the current smoker. That is $P(Y_x = 1|X; I_1 = 1; I_2 = 0) \neq P(Y_x = 1|X)$ and $P(Y_c = 1|X; I_1 = 1; I_2 = 1) \neq P(Y_c = 1|X)$. The direction of bias will be determined by the signs of the estimated selectivity coe±cients. The signs of the coe±cients (γ_{1x}; γ_{1c}) from the initiation equation are both positive, but they are not signi¯cant at the 5% level. On the other hand, the signs of coe±cients (γ_{2x}; γ_{2c}) from the cessation equation are both negative and signi¯cant at 5% level. Hence, they imply that the presence of the cessation selection causes an overestimation of the true risk factor for ex-smokers and an under estimation for the current smokers (see Eq. 4:1).

Our prediction shows that the true mean risk factor for the ex-smoker is about 6% which is slightly more than that of non-smokers and much less than

the observed risk factor (9.41%). The risk factor for the current smoker is about 20%, which is much higher than the observed risk factor of the current smoker (15.97%).[6] This finding has significant implications for previous empirical studies on the costs of cigarette smoking. For example, Miller et al. (1994) estimated medical care expenditures attributed to cigarette smoking based on observed frequencies of SRDs, without considering the unobserved heterogeneity in the baseline risk factors of smokers. In order to estimate the actual effect of cigarette smoking on health, this unobserved heterogeneity has to be first controlled for.

Further, our model is useful for investigating whether individual smoking participation decisions are consistent with economic rationality or forward looking behavior. The counter-factual conditional mean risk factor predictions are useful for this purpose, see Vella (1988). Table 7 presents the probabilities for getting the disease for nonsmokers, if they had chosen to be ex-smokers ($P[\hat{Y}_x = 1|X_n; I_1 = 0]$); and chosen to be current smokers ($P[\hat{Y}_c = 1|X_n; I_1 = 0]$); the probabilities for ex-smokers, if they had chosen not to start smoking ($P[\hat{Y}_n = 1|X_x; I_1 = 1; I_2 = 0]$), and had chosen not to quit smoking ($P[\hat{Y}_c = 1|X_x; I_1 = 1; I_2 = 0]$); and finally the probabilities for current smokers, if they had chosen not to start smoking ($P[\hat{Y}_n = 1|X_c; I_1 = 1; I_2 = 1]$), and if they had chosen to quit smoking ($P[\hat{Y}_x = 1|X_c; I_1 = 1; I_2 = 1]$).[7] However, $P[\hat{Y}_n = 1|X_c; I_1 = 1; I_2 = 1]$ and

[6] Note that the risk factors for the ever-smokers will be slightly higher than that of the non-smokers if we consider the differential sample attrition rates due to deaths for the three smoking groups. During the two-year period between waves 1 and 2, the annual mortality rates in our sample for the non-smokers, ex-smokers, and current smokers with SRDs (aged 52 and more) were calculated to be approximately 0.40%, 0.50%, and 1.50% respectively.

[7] Our conditional mean risk factor predictions are computed in the following way: $P[\hat{Y}_i = 1|X_j; I_1 = 1; I_2 = 1] = \frac{\Phi_3(X_j\beta_i; I_1^*; I_2^*; \rho_{1i}; \rho_{12}; \rho_{2i})}{\Phi_2(I_1^*; I_2^*; \rho_{12})}$ and similarly $P[\hat{Y}_i = 1|X_j; I_1 = 1] = \frac{\Phi_2(X_j\beta_i; I_1^*; \rho_{1i})}{\Phi_1(I_1^*)}$, where $i = n; x; c$ and $j = n; x; c$.

$P[\hat{Y}_n = 1|X_x; I_1 = 1; I_2 = 0]$ are not identifiable because our model can not identify γ_{2n}:

We find from Table 7 that the risk factor for an ex-smoker had he/she not chosen to quit ($P[\hat{Y}_c = 1|X_x; I_1 = 1; I_2 = 0]$) is even higher than that of a current smoker ($P[Y_c = 1|X_c; I_1 = 1; I_2 = 1]$). This evidence indicates that ex-smokers incorporate the hazards of smoking into their beliefs based on private information, and that they revert their smoking addiction because they could foresee health deterioration. Current smokers may not fully realize the hazards of smoking because they have not yet run down their health stock below an individual-specific critical threshold. Others may be ignoring the signs of health deterioration, or unable to quit simply because of addiction.

Another interesting aspect of our empirical model is that it can examine the presence and direction of comparative advantage (or more appropriately, \comparative risk" in our context) in the initiation and cessation decisions. Our empirical evidence has indicated that self-selection has a non-ignorable effect on the observed risk factors for both ex-smokers and the current smokers. Under self-selection, individuals will choose an alternative for which they have a comparative advantage (see Sattinger (1978), Fishe et al. (1981), and Maddala (1983)). The existence of the comparative risk will suggest a lack of forward looking behavior in that choice.

First, we examine the presence of comparative risk in the cessation decision by looking at the counter-factual conditional mean risk predictions, $P[\hat{Y}_n = 1|X_x; I_1 = 1; I_2 = 0]$ and $P[\hat{Y}_x = 1|X_c; I_1 = 1; I_2 = 1]$. In the absence of comparative risk in the cessation decision, we expect: $\frac{P[Y_x=1|X_x;I_1=1;I_2=0]}{P[\hat{Y}_x=1|X_c;I_1=1;I_2=1]} < \frac{P[\hat{Y}_c=1|X_x;I_1=1;I_2=0]}{P[Y_c=1|X_c;I_1=1;I_2=1]}$.

The left hand side of the inequality indicates the mean relative risk taken by observed quitters, whereas the right hand side of the inequality indicates the relative risk foregone by them. Based on the predictions presented in Table 7, we see that the inequality holds, which suggests rational risk-taking in the cessation decision.[8] Second, in order to examine the existence of comparative risk at the initiation stage, we obtain the following counter-factual mean risks : $P[\hat{Y}_n = 1 | X_x; I_1 = 1]$; $P[\hat{Y}_x = 1 | X_n; I_1 = 0]$; $P[\hat{Y}_x = 1 | X_x; I_1 = 1]$; $P[\hat{Y}_n = 1 | X_c; I_1 = 1]$; $P[\hat{Y}_c = 1 | X_n; I_1 = 0]$; and $P[\hat{Y}_c | X_c; I_1 = 1]$: Since we want to compare non-smokers with ex-smokers and current smokers in the initiation decision, the mean prediction here excludes the effect of the second selection. If the rational risk-taking behavior is valid on the average at the initiation stage, we expect the following inequalities to hold: $\frac{P[Y_n = 1 | X_n; I_1 = 0]}{P[\hat{Y}_n = 1 | X_x; I_1 = 1]} < \frac{P[\hat{Y}_x = 1 | X_n; I_1 = 0]}{P[\hat{Y}_x = 1 | X_x; I_1 = 1]}$; and $\frac{P[Y_n = 1 X_n; I_1 = 0]}{P[\hat{Y}_n = 1 | X_c; I_1 = 1]} < \frac{P[\hat{Y}_c = 1 | X_n; I_1 = 0]}{P[\hat{Y}_c = 1 | X_c; I_1 = 1]}$: Each side of the inequalities has similar interpretation as before. Our predictions show no evidence of forward looking behavior in the initiation decision.[9]

[8] The results from our predictions are:

$$\frac{P[Y_x = 1 | X_x; I_1 = 1; I_2 = 0]}{P[\hat{Y}_x = 1 | X_c; I_1 = 1; I_2 = 1]} = \frac{0:094}{0:082} = 1:146;$$

and

$$\frac{P[\hat{Y}_c = 1 | X_x; I_1 = 1; I_2 = 0]}{P[Y_c = 1 | X_c; I_1 = 1; I_2 = 1]} = \frac{0:214}{0:160} = 1:33:$$

[9] The results of our predictions are:

$$\frac{P[Y_n = 1 | X_n; I_1 = 0]}{P[\hat{Y}_n = 1 | X_x; I_1 = 1]} = \frac{0:054}{0:088} = 0:614;$$

$$\frac{P[\hat{Y}_x = 1 | X_n; I_1 = 0]}{P[Y_x = 1 | X_x; I_1 = 1]} = \frac{0:039}{0:067} = 0:582;$$

$$\frac{P[Y_n = 1 | X_n; I_1 = 0]}{P[\hat{Y}_n = 1 | X_c; I_1 = 1]} = \frac{0:054}{0:111} = 0:486;$$

These results are broadly consistent with those of Viscusi (1991) who found that young people have a high risk perception, but this risk perception does not influence their smoking behavior. Most at the initiation stage are too young to think about the hazard of smoking, which may or may not happen until many years in the future. At this age, they tend to experiment with various alternative life styles. The lack of rationality in the initiation behavior could be more transparent in our analysis because the hazards of cigarette smoking was not very well known to the public a few decades ago, and cigarette smoking was a more acceptable social behavior during the period in which the smokers in our sample initiated. Unlike the initiation decision, the rationality in the cessation decision is observed because individuals get first-hand information on the risk and utility of smoking from their past smoking experiences.

6. Conclusions

Almost all previous empirical studies on cigarette smoking estimated a particular part of our structural model and focused on the interpretation of the estimated coefficients. Thus a particular association between a dependent variable (smoking behavior or the presumed effect of smoking) and the covariates is the main issue in these studies. They uncover many interesting empirical regularities such as the demographic variations in smoking behavior, correlation of smoking with drinking habits, and so on. The estimation of our initiation and cessation equations corrob-

---

and

$$\frac{P[\hat{Y}_c = 1 | X_n; I_1 = 0]}{P[\hat{Y}_c = 1 | X_c; I_1 = 1]} = \frac{0:099}{0:205} = 0:483:$$

Since we are not conducting any statistical tests on the validity of these inequalities, these results are only suggestive.

orates results which are consistent with previous studies. However our objective is not to con¯rm these previous ¯ndings using new data, but to go beyond the mere interpretation of estimated regression coe±cients. Fundamentally, our two selection equations explore individual attitudes towards risk, and how these risk attitudes lead to di®erent health outcomes that is represented by the prevalence of SRDs. By combining smoking motivation and its outcome, our model reveals many useful aspects of how individuals incorporate their risk beliefs into smoking choices, and further it provides a clear direction for future public policy.

We ¯nd signi¯cant evidence of self-selection e®ects on both previous and current smokers' probabilities for getting SRDs. This indicates that previous studies on the e®ect of smoking on health or medical expenditures should be revisited after considering self-selection behavior of smokers. We ¯nd that the true mean risk factor for ex-smokers is about 6% which is slightly more than that of never-smokers, and much less than their observed risk factor (9.4%) in the sample. The true mean risk factor for the current smokers is about 20% which is much higher than their observed risk factor (16%). Based on counterfactual conditional mean risk factor predictions, we ¯nd that the risk factor for an ex-smoker, had he not chosen to quit, is even higher than that of a current smoker. Our evidence also suggests that self-selection in the cessation decision is consistent with economic rationality, but the same is not found in the initiation decision. This ¯nding underscores the importance of public policy initiatives on teenage initiation. In addition, our analysis suggests a direction for anti-smoking campaign; it should target those groups of individuals who tend to initiate more easily, because those who tend to initiate have higher risk factors for the diseases as well.

References

[1] Barsky, Robert B., F. T. Juster, M. S. Kimball, and M. D. Shapiro, Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study, The Quarterly Journal of Economics, May (1997), 537-579.

[2] Bartecchi, Carl E., T. D. MacKenzie, and R. W. Schrier, The human cost of tobacco use(¯rst of two parts), The New England Journal of Medicine 330 (13) (1994).

[3] Becker, Gary S. and Kevin Murphy, A theory of rational addiction, Journal of Political Economy 96 (4) (1988), 675-701.

[4] Breslau, Naomi, E. Peterson, Smoking cessation in young adults: Age at initiation of cigarette smoking and other suspected in°uences, American Journal of Public Health 86 (2) (1996a), 214-220.

[5] Breslau, Naomi, E. Peterson, L. Schultz,P. Andreski,and H. Chilcoat, are smokers with alcohol disorders less likely to quit?, American Journal of Public Health 86 (7) (1996b), 985-990.

[6] Blundell, R. and R. Smith, An exogeneity test for a simultaneous equation Tobit model with application to labor supply, Econometrica 54 (3) (1986), 679-685.

[7] Fielding, Jonathan E., Smoking: Health e®ects and control, The New England Journal of Medicine 313 (8) (1985), 491-497.

[8] Fishe, Raymond P. H., R. P. Trost, and P. M. Lurie, Labor force earnings and college choice of young women : An examination of selectivity bias and comparative advantage, Economics of Education Review, 1 (2) (1981) ,169-191.

[9] Go®e, W. L., G. Ferrier, and J. Rogers, Global optimization of statistical functions with simulated annealing, Journal of Econometrics 60 (1994), 65-99.

[10] Gourieroux, C., A. Monfort, E. Renault, and A. Trognon, Generalized residuals, Journal of Econometrics 34 (1987), 5-32.

[11] Greene, William H., Econometric Analysis, Macmillan, New York, 1993

[12] Hsieh, Chee-Ruey, Lee-Lan Yen, Jin-Tan Liu, and Chyongchiou Jeng Lin, Smoking ,health knowledge, and anti-smoking campaign: An empirical study in Taiwan, Journal of Health Economics 15 (1996), 87-104.

[13] Jones, Andrew M., Smoking cessation and health: a response, Journal of health Economics 15 (1996), 755-759.

[14] Jones, Andrew M., Health, addiction, social interaction and the decision to quit smoking, Journal of Health Economics 13 (1994), 93-110.

[15] Koop, Gary and D. J. Poirier, Learning about the across-regime correlation in switching regression models, Journal of Econometrics 78 (2) (1997), 217-227.

[16] Lahiri, Kajal and Jae G. Song, Testing for Normality in a Probit Model with Double Selection, Economics Letters, 65 (1999), 33-39.

[17] Leung, Siu Fai and Shihti Yu, On the Choice between Sample Selection and Two-part Models, Journal of Econometrics 72 (1996) 197-229.

[18] Maddala, G. S., Limited dependent and Qualitative Variables In Econometrics, Cambridge University press, Cambridge, 1983.

[19] Mattson, Margaret E., E. Pollack, and J. W. Cullen, What are the odds that smoking will kill you?, American Journal of Public Health 77 (4) (1987), 425-431.

[20] Miller, L. S., J. C. Bartlett, D. P. Rice,W. B. Max, and T. Novotny, Medical-Care Expenditures Attributable to Cigarette Smoking, 1993: Methodology, Descriptive Statistics, Parameter, and Expenditure Estimates, Center for Disease Control and Prevention, August, 1994.

[21] National Center for Health Statistics. Health, United States, 1995. Hyattsville, Maryland: Public Health Service, (1996), 173.

[22] Nawata, Kazumitsu and Ncbuco Nagase, Estimation of Sample Selection bias Models, Econometric Review, 1996 387-400.

[23] Nelson, Forrest D., E±ciency of the Two-Step Estimator for Models with Endogenous Sample Selection, Journal of Econometrics, 24 (1984), 181-196.

[24] Orphanides, Athanasios and David Zervos, Rational addiction with learning and regret, Journal of Political Economy 103 (4) (1995), 739-758.

[25] Pagan, Adrian and Frank Vella, Diagnostic tests for models based on individual data : A survey, Journal of Applied Econometrics 4 (1989), s29-s59.

[26] Sattinger, Michael, Comparative Advantage in individuals, The Review of Economics and Statistics, 60 (1978), 259-267.

[27] Shmueli, Amir, Smoking cessation and health: A comment, Journal of Health Economics 15 (1996), 751-754.

[28] Stolzenberg, Ross M., and Daniel A.Relles, Tools for intuition about sample selection bias ant its correction, American Sociology Review 65 (1997), 494-507.

[29] Vella, Francis, A simple estimator for simultaneous models with censored endogenous regressors, International Economic Review 34 (2) (1993), 441-457.

[30] Vella, Francis, Generating Conditional Expectations from Models with Selectivity Bias, Economics Letters 28 (1988), 97-103.

[31] Viscusi, W. Kip, Smoking: Making the Risky Decision, Oxford University Press, New York, 1992.

[32] Viscusi, W. Kip, Age variations in risk perceptions and smoking decisions, The Review of Economics and Statistics 73 (4) (1991), 577-589.

[33] Viscusi, W. Kip, Do smokers underestimate risks?, Journal of Political Economy 98 (6) (1990), 1253-1269.

[34] Yatchew, A., and Z. Griliches, Speci¯cation Error in Probit Models, Review of Economics and Statistics 66 (1984), 134-139.

[35] Yuan, Jian-Min, R. K. Ross, X. L. Wang, Y. T. Gao, B. E. Henderson, and M. C. Yu, Morbidity and Mortality in relation to cigarette smoking in Shanghai, China, Journal of American Medical Association 275 (21) (1996), 1646-1650.

Appendix A

Variable Names and Definitions

Variable	Definition
Except otherwise noted, all variables denoted below are (0 1) dummies.	
Initiation Equation	
Dependent variable	Ever-smoker = 1, never smoker = 0
BMI	Body Mass Index (weight/height^2) normalized by sample mean (not a dummy).
CATHOLIC	Religious preference, Catholic.
CLEAN	Interior of the dwelling units is clean.
EVDRINK	Ever drink any alcoholic beverages such as beer, wine, or liquor.
EVRSMOKD	Ever smoked cigarette.
EXDRINK	Excessive drinking habits.
FINJOB	Worked in the business of finance, insurance or real estate. (US Census Code :700-712)
FORNBORN	Not born in the U.S.
INVEST	Invest in stock, bond, real estate, or t-bill.
MALE	Male.
MARID1ST	First marriage and currently married.
MILIT	Ever been in military service.
MYOPIC	The most important financial planning horizon for the individual is next few months.
NETWORTH	Total net worth normalized by sample mean (not a dummy).
NVRWORKD	Never worked for pay.
PAREDU	Either mother or father has more than 12 years of schooling.
RACEB	Black.
RACEW	White/Caucasian.
RELIGS	Attend religious services more than two or three times a month.
RISKAVER	Will not take new job with a 50-50 chance of doubling income and 50-50 chance of cutting income by half.
SERVJOB	Service related job (US Census Code :403-407, 413-427, 433-469).
SALESJOB	Worked in the business of retail or whole sale (US Census Code: 500-571, 580-690).
SCHLYRS	Years of schooling normalized by sample mean (not a dummy).
SCHLYRS2	SCHLYRS^2
SOUTHB	Born in southern states.
SPOUSNSM	Individual's spouse is nonsmoker.
TECHJOB	Professional specialty operation and technical support job (US Census Code: 043-235).
WESTB	Born in western states.

Continuation Equation

Dependent variable	Current smoker = 1, ex-smoker = 0
BMI2	BMI^2
ADDICTION	Years of smoking times number of cigarette smoked, normalized by sample mean (not a dummy).
ALCHOLIC	Drink more than 3 or 4 a day.
ARTHRTS	Ever had arthritis.
BADFIN	Dissatisfied with his/her financial condition.
BLOODPRS	Ever had high blood pressure problem.
CHOLSTRL	Ever had high cholesterol problem.
DIABTS	Ever had diabetes.
EXER	Do both light and heavy exercise more than three time a week.
FINJOB	Worked in the business of finance, insurance, and real estate (US Census Code : 700-712).
HELTHINS	Individual has health insurance policy, either federally funded, privately funded, or employer provided.
HOUSE	Living in a detached single family house.
IRA	Individual has IRA account.
JOGAMILE	Individual has no difficulty at all in running or jogging a mile.
LIFEINS	Individual has one or more life insurance policy.
MARRIED	Currently married.
NOMARAGS	Number of marriages including current one normalized by sample. mean (not a dummy).
OWNRV	Own recreational vehicle(RV).
RACEB	Race dummy for Black.
SPOSEXSM	Individual's spouse is ex-smoker.
R12(σ_{12})	Inverse of Mill's ratio for the first selection equation(initiation).

BMI, CATHOLIC, CLEAN, FINJOB, FORNBORN, MALE, MYOPIC, RELIGS RISKAVER, SALESJOB, SCHLYRS, SCHLYRS2, SPOUSNSM have been already defined before.

Disease equations

Dependent variable	Ever had smoking related diseases =1, Never had the disease = 0. Smoking related diseases (SRDs): lung disease (not including asthma) such as chronic bronchitis or emphysema, or any following type of the cancers of the abdomen, mouth, bladder, neck, nose, Pancreas, bronchia, cervix, esophagus, stomach, throat, tongue, kidney, liver, and lung.
AGE	Age normalized by sample mean (not a dummy).
AGRIGJOB	Work in the area of agriculture, forestry, and fishing (U.S. Census Code: 010-031).

CIGARETS	Number of cigarette smoked per day normalized by sample mean.
SALESMAN	Sales job (U.S. Census Code: 243-285).
FEDINS	Federally funded health insurance, such as Medicaid, CHAMPUS, VA, or other military programs.
GOODFAM	Satisfied family life
HAZWORK	Ever exposed and continuously exposed more than 1 year to a dangerous chemicals or other hazards at work.
NOJOBS	Number of jobs including current one normalized by sample mean (not a dummy).
R12(σ_{12})	Inverse of Mill's ratio for the first selection equation (initiate), ever-smoker.
R1n(σ_{1n})	Inverse of Mill's ratio for the first selection equation (not initiate), nonsmoker.
R1x(σ_{1x})	Inverse of Mill's ratio for the first selection equation (initiate), ex-smoker
R2x(σ_{2x})	Inverse of Mill's ratio for the second selection equation (quit), ex-smoker.
R1c(σ_{1c})	Inverse of Mill's ratio for the second selection equation (initiate), current smoker.
R2c(σ_{2c})	Inverse of Mill's ratio for the second selection equation (continue), current smoker.

ARTHRITS, BADFIN, BLOODPRS, BMI, BMI2, CATHOLIC, CHOLSTRL, DIABTS, EXDRINK, FORNBORN, LIFEINS, MARID1ST, MYOPIC, NETWORTH, NEVERW NOMARAGS, RISKAVER, RACEB, SCHLYRS, SERVJOB, SPOSEXSM and SPOSENSM have already been defined above.

Table 1.

Descriptive Statistics for Selected Variables

variables	whole sample mean	nonsmoker mean	ex-smoker mean	current smoker mean
AGE	1.0000	0.9967	1.0088	0.9922
ALCHOLIC	0.0543	0.0179	0.0537	0.1039
ARTHRTS	0.3974	0.3836	0.4041	0.4067
BADFIN	0.2244	0.1926	0.1900	0.3142
BLOODPRS	0.4062	0.4061	0.4353	0.3663
BMI	1.0000	1.0116	1.0189	0.9586
CATHOLIC	0.2734	0.2616	0.2892	0.2673
CHOLSTRL	0.2401	0.2443	0.2553	0.2135
DIABTS	0.1123	0.1071	0.1232	0.1043
DISEASE	0.0974	0.0542	0.0941	0.1597
EVDRINK	0.6078	0.5196	0.6660	0.6459
EXDRINK	0.2031	0.0904	0.2524	0.2865
EXER	0.2098	0.2212	0.2432	0.1487
FEDINS	0.1671	0.1263	0.1859	0.1960
FINJOB	0.0520	0.0593	0.0454	0.0513
FORNBORN	0.0973	0.1238	0.0900	0.0717
GOODFAM	0.2211	0.2379	0.2396	0.1732
HAZWORK	0.2040	0.1552	0.2224	0.2441
HELTHINS	0.8572	0.8598	0.8893	0.8097
INVEST	0.6253	0.6660	0.6758	0.5016
IRA	0.4197	0.4551	0.4822	0.2865
JOGAMILE	0.1447	0.1618	0.1571	0.1047
LIFEINS	0.7098	0.6988	0.7527	0.6659
MALE	0.5003	0.3614	0.6161	0.5273
MARID1ST	0.5405	0.6109	0.5621	0.4165
MARIED	0.7508	0.7682	0.7975	0.6634
MILIT	0.2956	0.1935	0.3833	0.3121
MYOPIC	0.1829	0.1682	0.1609	0.2327
NETWORTH	1.0000	1.1744	1.0607	0.6834
NEVERW	0.0349	0.0554	0.0187	0.0297
PAREDU	0.2999	0.2823	0.3141	0.3040
RACEB	0.1634	0.1667	0.1443	0.1850
RACEW	0.7254	0.7037	0.7548	0.7143
RELIGS	0.5256	0.6413	0.5220	0.3757
RISKAVER	0.8717	0.8862	0.8735	0.8496
SALESJOB	0.1547	0.1363	0.1485	0.1879
SCHLYRS	1.0000	1.0202	1.0171	0.9494
SERVJOB	0.0924	0.0931	0.0852	0.1015
SPOSENSM	0.5117	0.5634	0.4840	0.4804
SPOSEXSM	0.2855	0.2878	0.3480	0.1968
TECHJOB	0.1277	0.1451	0.1455	0.0799
WESTB	0.0784	0.0803	0.0808	0.0725

Table 2.

Estimates for the Initiation Equation

variable	two-step	FIML	S. E. (FIML)	odds ratio	marginal effect
Constant	0.8821	0.8964	0.1599	2.4508	0.3091
MALE	0.4142	0.4173	0.0451	1.5179	0.1452
RACEW	0.0871	0.0835	0.0559	1.0871	0.0305
RACEB	0.1053	0.1101	0.0754	1.1164	0.0369
FORNBORN	-0.2378	-0.2483	0.0591	0.7801	-0.0833
WESTB	-0.1118	-0.1104	0.0554	0.8955	-0.0392
SOUTHB	0.0189	0.0145	0.0339	1.0146	0.0066
SCHLYRS	0.5224	0.5205	0.2507	1.6829	0.1831
SCHLYRS2	-0.5048	-0.5071	0.1397	0.6022	-0.1769
PAREDU	0.0740	0.0784	0.0308	1.0816	0.0259
CATHOLIC	0.1495	0.1520	0.0362	1.1642	0.0524
RELIGS	-0.2729	-0.2764	0.0341	0.7585	-0.0956
MILIT	0.1477	0.1487	0.0424	1.1603	0.0518
MARID1ST	-0.3707	-0.3710	0.0354	0.6900	-0.1299
EVDRINK	0.2950	0.2996	0.0423	1.3493	0.1034
EXDRINK	0.8802	0.8823	0.1904	2.4165	0.3085
SPOUSNSM	-0.3499	-0.3488	0.0343	0.7055	-0.1226
NETWORTH	-0.0155	-0.0156	0.0057	0.9845	-0.0054
INVEST	-0.1482	-0.1489	0.0354	0.8617	-0.0519
SALESJOB	0.1236	0.1270	0.0397	1.1354	0.0433
FINJOB	-0.0698	-0.0691	0.0607	0.9332	-0.0245
TECHJOB	0.0787	0.0798	0.0477	1.0831	0.0276
SERVJOB	0.0408	0.0401	0.0469	1.0409	0.0143
NVRWORKD	-0.2655	-0.2677	0.0778	0.7651	-0.0931
RISKAVER	-0.0817	-0.0826	0.0428	0.9207	-0.0286
MYOPIC	0.0650	0.0678	0.0384	1.0702	0.0228
CLEAN	-0.0763	-0.0773	0.0292	0.9256	-0.0267
BMI	-0.3763	-0.3834	0.0820	0.6815	-0.1319
RACEB	0.3005	0.3088	0.0826		
MARID1ST	-0.2341	-0.2302	0.0602		
EVDRINK	0.1838	0.1862	0.0586	*Heteroskedasticity*	
EXDRINK	0.3354	0.3348	0.1334	*Specification*	
SPOUSNSM	-0.1738	-0.1677	0.0561		

Table 3. Estimates for the continuation equation

variable	two-step	FIML	S. E. (FIML)	odds ratio	marginal effect
Constant	0.9862	1.0069	0.2036	2.7371	0.3900
MALE	-0.1056	-0.1202	0.0346	0.8867	-0.0417
RACEB	0.1491	0.1619	0.0701	1.1757	0.0590
RACEW	0.0917	0.1343	0.0662	1.1437	0.0363
FORNBORN	-0.1211	-0.1094	0.0534	0.8964	-0.0479
CATHOLIC	0.0556	0.0489	0.0264	1.0501	0.0220
RELIGS	-0.1972	-0.1968	0.0351	0.8214	-0.0779
MARRIED	-0.1635	-0.1633	0.0451	0.8493	-0.0646
NOMARAGS	0.0450	0.0443	0.0197	1.0453	0.0178
HOUSE	-0.0533	-0.0497	0.0250	0.9515	-0.0211
EXER	-0.1408	-0.1485	0.0341	0.8620	-0.0557
ALCHOLIC	0.1896	0.1858	0.0495	1.2042	0.0749
CLEAN	-0.1100	-0.1156	0.0267	0.8908	-0.0435
BADFIN	0.0799	0.0814	0.0287	1.0848	0.0316
SPOUSNSM	-0.2940	-0.2973	0.0574	0.7428	-0.1163
SPOUSXSM	-0.4094	-0.4209	0.0600	0.6565	-0.1619
OWNRV	-0.1512	-0.1566	0.0654	0.8550	-0.0598
IRA	-0.1353	-0.1465	0.0309	0.8637	-0.0535
LIFEINS	-0.0373	-0.0408	0.0260	0.9600	-0.0148
RISKAVER	-0.0498	-0.0486	0.0319	0.9526	-0.0197
HELTHINS	-0.0872	-0.0827	0.0339	0.9206	-0.0345
SCHLYRS	0.5217	0.5413	0.2149	1.7182	0.2063
SCHLYRS2	-0.3956	-0.4102	0.1186	0.6635	-0.1564
JOGAMILE	-0.1466	0.1696	0.0417	1.1848	-0.0580
BLOODPRS	-0.0946	-0.0983	0.0266	0.9064	-0.0374
CHOLSTRL	-0.0421	-0.0443	0.0254	0.9567	-0.0166
DIABTS	-0.0356	-0.0385	0.0355	0.9622	-0.0141
ARTHRTS	-0.0260	-0.0280	0.0226	0.9724	-0.0103
SALESJOB	0.0745	0.0725	0.0294	1.0752	0.0295
FINJOB	0.0946	0.0978	0.0492	1.1027	0.0374
MYOPIC	0.0691	0.0700	0.0296	1.0725	0.0273
ADDICTION	0.0595	0.0593	0.0142	1.0611	0.0235
BMI	-0.7590	-0.8134	0.1165	0.4433	-0.3001
R12	0.1137	0.1590	0.1227		
MALE	0.1671	0.1516	0.0671		
RACEW	-0.5697	-0.5365	0.0904		
MARRIED	-0.1372	-0.1209	0.0951	*Heteroskedasticity*	
SPOUSNSM	-0.2414	-0.2112	0.0832	*Specification*	
OWNRV	0.2249	0.1994	0.1233		

Table 4.

Estimates for the Nonsmoker's Disease Equation

variable	two-step	FIML	S. E. (FIML)	odds ratio	marginal effect
Constant	-2.7242	-2.6669	0.9493	0.0695	-0.3230
MALE	-0.0873	-0.0747	0.1033	0.9280	-0.0104
RACEB	0.2413	0.2329	0.1699	1.2623	0.0286
FORNBORN	0.4258	0.4149	0.1692	1.5142	0.0505
WESTB	0.6618	0.6488	0.2103	1.9132	0.0785
MARID1ST	-0.0532	-0.0605	0.0898	0.9413	-0.0063
BLOODPRS	0.4400	0.4388	0.1589	1.5508	0.0522
CHOLSTRL	0.1947	0.1931	0.0760	1.2130	0.0231
DIABTS	-1.0859	-1.0799	0.5215	0.3396	-0.1287
ARTHRTS	0.2126	0.2132	0.0787	1.2376	0.0252
EXDRINK	-0.1684	-0.1548	0.1285	0.8566	-0.0200
BADFIN	0.2037	0.2042	0.0832	1.2265	0.0242
GOODFAM	0.0344	0.0324	0.1030	1.0329	0.0041
HAZWORK	0.3111	0.3122	0.0871	1.3664	0.0369
AGRIJOB	0.1120	0.1108	0.1879	1.1172	0.0133
SALESMAN	-0.0824	-0.0806	0.1495	0.9226	-0.0098
SERVJOB	-0.0907	-0.0885	0.1188	0.9153	-0.0108
NVRWORKD	0.5355	0.5319	0.1786	1.7022	0.0635
FEDINS	0.2415	0.2406	0.0933	1.2720	0.0286
LIFEINS	0.0312	0.0306	0.0808	1.0311	0.0037
RISKAVER	-0.0051	-0.0081	0.1106	0.9919	-0.0006
MYOPIC	-0.2119	-0.2067	0.1512	0.8133	-0.0251
SPOUSNSM	0.1722	0.1634	0.1291	1.1775	0.0204
SPOUSXSM	0.1780	0.1773	0.1281	1.1940	0.0211
NOJOBS	-0.0264	-0.0256	0.0641	0.9747	-0.0031
NOMARAGS	0.1876	0.1884	0.0863	1.2073	0.0222
AGE	0.3834	0.3763	0.4647	1.4569	0.0455
SCHOLYRS	-0.0697	-0.0741	0.1334	0.9286	-0.0083
NETWORTH	-0.0426	-0.0433	0.0344	0.9576	-0.0051
BMI	0.1914	0.1863	1.2774	1.2048	0.0227
BMI2	0.0138	0.0144	0.5548	1.0145	0.0016
R1n	-0.0247	0.0130	0.1791		
RACEB	-0.3961	-0.3923	0.1701		
FORNBORN	-0.5785	-0.5774	0.1665		
WESTB	-0.6623	-0.6529	0.2747		
BLOODPRS	-0.3126	-0.3127	0.1420	*Heteroskedasticity*	
DIABTS	0.9055	0.9067	0.2777	*Specification*	
NVRWORKD	-0.7137	-0.7206	0.2851		
MYOPIC	0.4190	0.4189	0.1356		

Table 5.

Estimates for the Ex-smoker's Disease Equation

variable	two-step	FIML	S. E. (FIML)	odds ratio	marginal effect
Constant	-0.8954	-1.1841	0.8318	0.3060	-0.1379
MALE	-0.1548	-0.1390	0.0948	0.8702	-0.0238
RACEB	-0.1923	-0.1884	0.1025	0.8283	-0.0296
FORNBORN	-0.5729	-0.5346	0.2005	0.5859	-0.0882
WESTB	-0.0830	-0.0777	0.1117	0.9252	-0.0128
MARID1ST	0.0794	0.0886	0.0911	1.0926	0.0122
BLOODPRS	-0.2342	-0.3002	0.1871	0.7407	-0.0361
CHOLSTRL	0.0018	0.0083	0.0692	1.0083	0.0003
DIABTS	0.1398	0.1355	0.0886	1.1451	0.0215
ARTHRTS	0.1581	0.1585	0.0647	1.1718	0.0244
EXDRINK	0.1066	0.0986	0.0784	1.1036	0.0164
BADFIN	0.0811	0.0602	0.0777	1.0620	0.0125
GOODFAM	0.2763	0.3265	0.1623	1.3861	0.0425
HAZWORK	0.4941	0.5138	0.1582	1.6716	0.0761
AGRIJOB	-0.0056	-0.0026	0.1825	0.9974	-0.0009
SALESMAN	-0.9838	-1.3988	1.3063	0.2469	-0.1515
SERVJOB	-0.0651	-0.0654	0.0962	0.9367	-0.0100
NVRWORKD	-0.2946	-0.2985	0.2684	0.7419	-0.0454
FEDINS	0.2136	0.2146	0.0838	1.2394	0.0329
LIFEINS	-0.0581	-0.0461	0.0698	0.9549	-0.0090
RISKAVER	-0.1243	-0.1189	0.0910	0.8879	-0.0191
MYOPIC	0.2735	0.3380	0.1666	1.4021	0.0421
SPOUSNSM	0.0467	0.0650	0.1030	1.0672	0.0072
SPOUSXSM	0.1109	0.1459	0.1039	1.1571	0.0171
NOJOBS	-0.0631	-0.0683	0.0455	0.9340	-0.0097
NOMARAGS	0.1436	0.1392	0.0681	1.1494	0.0221
AGE	0.3400	0.3822	0.4096	1.4655	0.0524
SCHOLYRS	-0.3158	-0.2746	0.1272	0.7599	-0.0486
NETWORTH	-0.0209	-0.0196	0.0217	0.9806	-0.0032
BMI	-1.8797	-1.6930	1.2346	0.1840	-0.2895
BMI2	0.8676	0.8014	0.5460	2.2287	0.1336
CIGARETS	0.2299	0.2219	0.0448	1.2484	0.0354
R1x	0.2359	0.2230	0.2485		
R2x	-0.2189	-0.3908	0.1743		
BLOODPRS	0.2710	0.2659	0.1206		
GOODFAM	-0.4824	-0.4197	0.1386		
HAZWORK	-0.3551	-0.2886	0.1475	*Heteroskedasticity*	
SALESMAN	0.6186	0.6660	0.4368	*Specification*	
MYOPIC	-0.3588	-0.3315	0.1439		

Table 6.

Estimates for the Current Smoker's Disease Equation

variable	two-step	FIML	S. E. (FIML)	odds ratio	marginal effect
Constant	-0.9730	-0.8839	0.8575	0.4132	-0.1590
MALE	0.0275	-0.0065	0.1638	0.9935	0.0045
RACEB	-0.5621	-0.5266	0.1262	0.5906	-0.0919
FORNBORN	-0.6234	-0.6189	0.1976	0.5385	-0.1019
WESTB	-0.2598	-0.2481	0.1523	0.7803	-0.0425
MARID1ST	0.0584	0.0569	0.0976	1.0585	0.0095
BLOODPRS	0.1616	0.1477	0.0811	1.1592	0.0264
CHOLSTRL	0.1864	0.1697	0.0915	1.1849	0.0305
DIABTS	-0.1212	-0.0644	0.2057	0.9376	-0.0198
ARTHRTS	0.3358	0.3250	0.0798	1.3840	0.0549
EXDRINK	0.0223	0.0043	0.1047	1.0043	0.0036
BADFIN	0.2615	0.2610	0.0951	1.2982	0.0427
GOODFAM	-0.2212	-0.2172	0.1306	0.8048	-0.0361
HAZWORK	0.2300	0.2224	0.0935	1.2491	0.0376
AGRIJOB	0.5055	0.4835	0.2286	1.6217	0.0826
SALESMAN	-0.3303	-0.3215	0.1484	0.7251	-0.0540
SERVJOB	-0.1556	-0.1485	0.1196	0.8620	-0.0254
NVRWORKD	0.5811	0.5481	0.1791	1.7300	0.0950
FEDINS	0.4773	0.4648	0.1025	1.5917	0.0780
LIFEINS	-0.2028	-0.1435	0.1442	0.8663	-0.0331
RISKAVER	-0.0055	-0.0086	0.1035	0.9914	-0.0009
MYOPIC	0.0855	0.0894	0.0869	1.0935	0.0140
SPOUSNSM	0.1659	0.1583	0.1065	1.1715	0.0271
SPOUSXSM	0.2418	0.2105	0.1355	1.2343	0.0395
NOJOBS	0.0061	0.0026	0.0586	1.0026	0.0010
NOMARAGS	0.1321	0.1352	0.0590	1.1448	0.0216
AGE	1.0138	0.9496	0.5889	2.5847	0.1657
SCHOLYRS	-0.2571	-0.2513	0.1692	0.7778	-0.0420
NETWORTH	-0.0249	-0.0237	0.0414	0.9766	-0.0041
BMI	-2.2286	-2.2623	1.0911	0.1041	-0.3642
BMI2	0.6494	0.6637	0.5072	1.9420	0.1061
CIGARETS	0.1854	0.1875	0.0682	1.2062	0.0303
R1c	0.1777	0.0931	0.3218		
R2c	-0.3443	-0.2373	0.1452		
MALE	-0.1564	-0.1497	0.1336		
DIABTS	0.3646	0.3744	0.2216	*Heteroskedasticity*	
NVRWORKD	-0.8484	-0.9518	0.3737	*Specification*	
LIFEINS	0.3288	0.3080	0.1316		

Table 7

Counter-factual Risk Factor Predictions

Observed Smoking Status	*Counter-factual Smoking Status*		
	$E[Y_n \mid .]$ (None)	$E[Y_X \mid .]$ (Ex.)	$E[Y_C \mid .]$ (Current)
None (I_1 =0)	0.054	0.056	0.111
Ex (I_1 =1, I_2 =0)	*	0.094	0.214
Current (I_1 =1, I_2 =1)	*	0.082	0.159