A Dynamic Stochastic Model of Lifetime Smoking Behavior

Michael Darden

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Economics.

Chapel Hill
2011

Approved by:

Donna Gilleskie, Advisor

R. Vijay Krishna

Brian McManus

Frank Sloan

Helen Tauchen

Abstract

MICHAEL DARDEN: A Dynamic Stochastic Model of Lifetime Smoking Behavior.
(Under the direction of Donna Gilleskie.)

This dissertation discusses results obtained through formulation and estimation of a dynamic stochastic model that captures individual smoking decision making, health expectations, and longevity over the life cycle. The standard rational addiction model is augmented with a Bayesian learning process about the health marker transition technology to evaluate the importance of personalized health information in the decision to smoke cigarettes. Additionally, the model is well positioned to assess how smoking and smoking cessation impact morbidity and mortality outcomes while taking into consideration the potential for dynamic selection of smoking behaviors. This research also provides a novel approach to the empirical construction of the theoretically common "smoking stock" that facilitates the estimation of investment and depreciation parameters. The structural parameters are estimated using rich longitudinal health and smoking data from the Framingham Heart Study: Offspring Cohort. Results suggest that there exists heterogeneity across individuals in the pathways by which smoking effects health. Furthermore, upon smoking, the estimated parameters suggest a positive reinforcement effect and a negative withdrawal effect, both of which encourage future smoking. I find that only in the case of very large change in an individual's health markers will the associated change in beliefs induce individuals to quit smoking. Generally, personalized health marker information is not found to influence smoking behavior relative to chronic health shocks themselves. The dissertation also presents evidence of health selection in smoking behavior that, when not modeled, may cause an overstatement of the direct effect of smoking on morbidity and mortality.

Acknowledgments

This dissertation would not have been possible without the help and guidance of several people. First, the support of my family has been instrumental in my success. Thank you, Mom, Dad, Jesse, and Lucas for all of your love and encouragement. Next, I must thank Donna Gilleskie for all of her hard work and dedication in training me for a career in economic research. I could not have asked for a better dissertation advisor. I would also like to thank the members of my dissertation committee, Helen Tauchen, Brian McManus, Vijay Krishna, and Frank Sloan, for the array of helpful comments that they have provided. Finally, I thank Manasigan Kanchanachitra for her constant support.

The Framingham Offspring Study (FOS) is conducted and supported by the NHLBI in collaboration with the FOS Study Investigators. This manuscript was prepared using a limited access dataset obtained from the NHLBI and does not necessarily reflect the opinions or views of the FOS or the NHLBI.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The decision to smoke has long interested social scientists and health policy researchers because of the seemingly irrational nature of such a choice. Why would an individual undertake an activity with such clear negative health consequences? A thorough review of this debate can be found in Sloan, Smith, and Taylor (2003). Ultimately, those authors conclude that individuals make decisions within an environment that reflects individual preferences but one that is also subject to information acquisition costs. Gary Becker describes economic decision makers: "They (economic agents) are not expected to be perfect optimizers, as evaluated by the analyst, or dispassionate external observers; rather, people do the best they can, given their information and their cognitive abilities to understand it (qtd. in Sloan *et al.* (2003) pg. 25)." An important question addressed by the smoking literature has been: what determines and shapes "their information?" Furthermore, how does information influence smoking behavior? And, to what extent has information regarding the health effects of smoking been free from selection bias? These questions form the basis for this dissertation.

The purpose of this work is to analyze the relationship between the consumption of cigarettes and health in a dynamic discrete choice framework that incorporates learning. I estimate the structural parameters of an individual's optimization problem with the following trade-off: current enjoyment of cigarette consumption versus the associated uncertain future utility and health consequences. I consider two dimensions of health: health markers and chronic health. Health markers are those factors (e.g., blood pressure, cholesterol, etc.) viewed by the medical literature to significantly predict the onset of chronic conditions (e.g.,

cardiovascular disease, cancer, etc.).[1] Given a history of these health markers and smoking behaviors, an individual is able to more precisely evaluate the effect of smoking on her health markers levels which, in turn, helps to determine her chronic health probability. More generally, health markers offer information as to an individual's overall health condition. Endowed with this information, an individual makes the smoking choice that maximizes her present discounted expected utility. Smoking history is modeled as a capital stock and is measured in a novel way so as to facilitate the estimation of depreciation and investment coefficients while keeping the model computationally tractable.

The structural parameters of the model are estimated with rich longitudinal data from the offspring of the original cohort of the Framingham Heart Study. I estimate preference parameters jointly with smoking stock, health marker, chronic health, and mortality transition parameters. Additionally, I estimate a flexible random effects specification that captures the correlation in permanent unobserved heterogeneity across outcomes. I use a value function interpolation method and simulated maximum likelihood to solve and estimate the model.

This dissertation fits into and extends the literature in four ways. First, the structural model extends the standard rational addiction model of Becker and Murphy (1988) by incorporating both health and learning. I model individuals as forward-looking in the sense that they evaluate current smoking alternatives while taking into consideration the future health and utility consequences associated with past and current smoking behavior. Furthermore, while other papers that incorporate health into the rational addiction framework (e.g., Suranovic *et al.* (1999); Carbone *et al.* (2005)) are purely theoretical, I *estimate* the preference and expectation parameters of a rational addiction model that capture forward-looking behavior with respect to health. Consistent with the theory of rational addiction, I find that smoking is reinforcing in the sense that the marginal utility of smoking is increasing in the amount of past smoking. I also find that the costs of withdrawal can prevent individuals from quitting smoking. The reinforcement and withdrawal effects are found to drive smoking dynamics.

[1] See Chapter 4 for a complete list of the health markers and chronic health conditions considered.

Second, a major contribution of this dissertation is to model how the receipt of information about an individual's health markers may alter an individual's smoking behavior. When information about the ill-health effects of smoking is made personal, the literature has argued that this personalized health information may provide an individual with a powerful incentive or "wake-up call" to curtail her smoking behavior (Sloan *et al.*, 2003). As an example, if a smoker experiences a heart attack, this health shock may be framed as an informative signal of the consequences of her smoking.[2] This dissertation extends the literature on smoking responses to personalized health information that have only considered chronic health shocks (e.g., heart attacks, cancer diagnoses, etc.) (Smith *et al.*, 2001; Khwaja *et al.*, 2006; Arcidiacono *et al.*, 2007). The distinction between health marker information (i.e., observing one's own blood pressure at a health exam) and chronic health shocks *as information* is important if the potential gains from information from a chronic health shock are "too late". Indeed, if policy makers are concerned with improving overall expected longevity, the incidence of a heart attack, while effective in convincing an individual to quit smoking, may not yield additional longevity. However, if changes in blood pressure, say, over time that are due to smoking, convince an individual to quit prior to the incidence of a heart attack, then there may exist a role for policy to emphasize the personalized "warning signs." Thus, motivated by the Becker quote above, my model explicitly accounts for health marker transition learning and I estimate parameters that dictate a Bayesian learning process.

To evaluate the roles of learning and information, I use the model and the estimated structural parameters to simulate smoking behavior and health and mortality outcomes under different counterfactual scenarios. The results suggest that there exists heterogeneity across individuals in the pathways by which smoking affects health. I find that the effect of an accumulated smoking stock on health markers varies widely across individuals relative to the mean effect. While the average variance in beliefs regarding this effect decreases by 20% after the

[2]Specifically, Sloan *et al.* (2003) and Khwaja *et al.* (2006) argue that, while general sources of health information (e.g., commercials, warning labels, etc.) may influence the light to moderate smoker, these sources of information will have less affect on the smoking behavior of heavy smokers. Indeed, for the heavy smoker, personalized information is required to change behavior. I will discuss the distinction between general and personalized health information as it relates to smoking behavior in Chapter 2.

first health exam, the estimated mean of the parameter distribution is small and thus, does not greatly impact smoking behavior. Counterfactuals in which individuals receive information more and less frequently do not appear to show individuals altering their smoking behavior in any significant sense. Indeed, only when an individual observes a discrete shock to the 90th percentile of the health marker distribution does she begin to smoke less. While smokers are found to quit or reduce their smoking after a chronic health shock, results from this dissertation suggest that health markers have limited informational value. Chapter 8 provides a detailed description of the results from these counterfactual simulations.

Third, using data from across the life cycle, I measure the role of health and mortality transition determinants by estimating these production technologies within the structural model of lifetime smoking decisions. This method allows both for smoking to affect health and for rational individuals to select the optimal smoking alternative while taking into account their health. My method improves upon recent papers that estimate health transitions outside the structural model (Adda and Lechene, 2001). Modeling both smoking behavior and health outcomes also allows me to capture unobserved heterogeneity that may be present in their joint determination. For example, if smokers are also more likely to engage in other health-hazardous behaviors (e.g. drinking, drug use, sky diving, etc.), then treating smoking as random within the population will lead to overstated estimates of the direct effect of smoking on the health outcome of interest. Importantly, alleviating this bias does not require the researcher to model every potential lifestyle choice that may be correlated with both smoking and health outcomes. Rather, to capture unobserved heterogeneity, I allow the unobserved errors that affect smoking, health, and mortality to be serially correlated through a common permanent unobserved component. I model this error structure with the discrete factor method (Heckman and Singer, 1984; Mroz, 1999) which amounts to a random effects specification of unobserved heterogeneity that is free from distributional assumptions. The error structure is similar to recent structural models that have accounted for unobserved heterogeneity (Arcidiacono *et al.*, 2007; Blau and Gilleskie, 2008).

To assess both the potential for selection on smoking and the effect of smoking on health outcomes, I simulate the structural model under different lifetime smoking patterns. For daily

light and heavy smoking from age 18, individuals can expect roughly 4.5 and 8 fewer years of longevity, respectively. These results are less severe in their overall assessment of the health effects of smoking on mortality than are the unconditional results presented in (Doll *et al.*, 1994, 2004).[3] My results are the first of which I am aware to explicitly control for the potential positive selection between smoking and health outcomes while also estimating the health transition equations within the structural model of lifetime smoking. Still, consistent with some literature on smoking and health[4], my results suggest that there exist longevity gains from quitting at any age, and that quitting prior to age 30 implies that an individual has roughly the same expected longevity as a lifelong nonsmoker. Furthermore, I find that quitting heavy smoking at ages 30, 40, 50, and 60 years of age increases life-expectancy by approximately 8, 7.75, 7, and 5.5 years, respectively. Interestingly, while I find that unobserved heterogeneity plays a major role in the dynamic relationship between smoking behavior and mortality, the unobserved heterogeneity plays almost no role in predicting health marker and chronic health transitions. My results indicate that there exists a strong positive correlation between smoking tendencies and underlying factors that influence mortality outcomes.[5]

Fourth, my model extends the empirical smoking literature with a novel construction of the "smoking stock". The key term in the standard rational addiction model of Becker and Murphy (1988) is an addictive capital stock that is subject to investment and depreciation. Those authors argue that addiction is captured by a positive interaction between the smoking stock and the marginal utility of addictive consumption. While theoretically it is quite intuitive to think an individual that has smoked in the past may have a higher marginal utility of smoking than someone who has not smoked, it is not clear how to best capture this capital stock empirically. For example, defining the stock to be the total number of years smoked ignores the

[3]See chapter 2 for a discussion on the medical view of smoking.

[4]See Doll *et al.* (1994); Taylor *et al.* (2002); Doll *et al.* (2004); Brønnum-Hansen *et al.* (2007)

[5]There is a large literature on the effects of smoking cessation on health outcomes. See United States Department of Health and Human Services (1990). This issue of "smoking depreciation" is often framed as the amount of time required after cessation until the probability of a health outcome converges to that of a nonsmoker. Because of the generality of my treatment of health (e.g., chronic health or not), and because the smoking cessation improves various health outcome probabilities differently, the discussion in this dissertation focuses on the benefits from quitting smoking on expected longevity.

importance of cessation (i.e., someone who smoked for 10 years and quit twenty years ago may have a different marginal utility for smoking than someone who smoked for 10 years and quit last year). Using factor analysis in a method similar to Sickles and Williams (2008), I create a continuous smoking stock index from several variables that reflect past smoking behavior.[6] This easily interpretable state variable captures the unique smoking history that each individual brings into each decision making period. Furthermore, measuring the smoking stock using this method also allows for the estimation of depreciation and investment parameters (Adda and Lechene, 2004).[7]

Finally, the structural model of this dissertation is solved and estimated using techniques common in the structural dynamic discrete choice modeling literature (Rust, 1987; Keane and Wolpin, 1994; Aguirregabiria and Mira, 2010). To incorporate learning, I combine several features of other recent structural papers that have explicitly modeled and estimated Bayesian learning processes (Ackerberg, 2003; Crawford and Shum, 2005; Chan and Hamilton, 2006; Mira, 2007; Chernew *et al.*, 2008). While those papers model learning about an intercept shifting term, I model learning about a marginal effect: the effect of the accumulated smoking stock on a set of health markers. Conditional on the unobserved heterogeneity, I use the model to predict the initial conditions (Khwaja, 2010).

This dissertation proceeds in the following chapters. Chapter 2 provides background on the economics and medical literatures with respect to smoking. That chapter begins with a discussion of the ubiquitous rational addiction model, including both theoretical extensions and alternatives and empirical applications. I also discuss the economics of information and expectations in the context of cigarette smoking. Finally, with respect to the medical literature on smoking, Chapter 2 places an emphasis on the effects of smoking on expected longevity.

[6] Sickles and Williams (2008) use principal component analysis to construct a "stock" of social capital from several correlated variables such as labor force participation, yearly hours worked, marriage, and "hours in income generating crime per year." See Chapter 2 for further details.

[7] Adda and Lechene (2004) note that the smoking stock may be different between reinforcing further smoking and affecting health. Indeed, those authors model two smoking stocks: a utility smoking stock and a health smoking stock. Each are subject to different investment and depreciation parameters that are fixed. In my case, I abstract from this possibility and estimate the transition parameters of a single stock variable that captures past smoking.

Chapter 3 presents the formal structural model. Chapter 4 discusses the Framingham Heart Study and presents basic summary statistics of the data used in estimation. How the structural model in Chapter 3 and the data in Chapter 4 are reconciled is discussed in Chapter 5. I also formally describe the econometric methods used in estimating the structural model as well as identification issues in Chapter 5. Parameter estimates and a brief discussion of how the estimated model fits the data are discussed in Chapter 6. In Chapter 7, I explain how I conduct the simulations that yield the main results of this dissertation. While counterfactual simulations that examine learning and information are present in Chapter 8, simulations that assess the effect of smoking on health are presented in Chapter 9. Finally, Chapter 10 concludes. Additionally, Appendix A gives the formal derivation of posterior beliefs.

Chapter 2

Background

Cigarette smoking is the single greatest preventable risk factor for mortality and morbidity. Tobacco smoke contains more that 7,000 chemicals. These chemicals quickly travel from the lungs into the blood stream, being distributed to *every* organ in the human body. Cigarette smoke increases both heart rate and blood pressure because the nicotine present in tobacco smoke activates the sympathetic nervous system (Bennett and Richardson, 1984; Omvik, 1996; Benowitz, 2003) (USDHHS, 2010, Chap. 6, pg. 368). Furthermore, medical evidence suggests that cigarette smoking both causes plaque to buildup in the arteries, causing an increase in total cholesterol, as well as, a decrease in high-density lipoprotein (hdl) "good" cholesterol (Garrison *et al.*, 1978) (USDHHS, 2010, Chap. 6, pg. 379-380). Smoking is even considered a risk factor for type 2 diabetes because cigarette smoke increases the degree of insulin resistance. (USDHHS, 2010, Chap. 6, pg. 383).

Cigarette smoking is causally linked to cancers of the bladder, cervix, esophagus, kidney, larynx, lung, mouth, pancreas, and stomach. Indeed, 9 out of 10 men who die from lung cancer have a smoking history. Furthermore, there exists a causal relationship between smoking and cardiovascular disease (both directly and indirectly through the effect of smoking on health markers such as blood pressure), atherosclerosis, various respiratory diseases, and several reproductive maladies. 440,000 deaths, nearly one-fifth of the annual deaths, are attributed to smoking in the United States each year. Illness from smoking is estimated to add $157 billion per year to national health expenditures, both from medical costs and lost productivity due to chronic illness and premature death (USDHHS, 2010). The 2004 United States Surgeon

General report on smoking concludes by stating: "Smoking harms nearly every organ of the body, causing many diseases and reducing the health of smokers in general." USDHHS, 2004

2.1 The Economics of Addiction

The rational addiction framework of Becker and Murphy (1988) has been the workhorse economic model of smoking behavior since its publication. The model is based on the premise that smokers and nonsmokers alike know the full price of smoking. That is, smokers are modeled as the forward-looking, utility maximizing agents of standard welfare economics. The key element of this framework is an accumulated experience of smoking term that is defined as a capital stock. By choosing to smoke in the current period, agents in the model realize that they are investing in their stock of smoking capital. The key dynamic in this model is that, by increasing the stock of smoking, an individual is altering the marginal utility of smoking in future periods. When the marginal utility of consumption of a particular good is increasing in that good's accumulated stock, Becker and Murphy (1988) categorize the good as addictive and exhibiting a complementary nature over time or "adjacent complementarity." The model is able to predict common observations such as binge behavior and "cold turkey" quitting.

Naturally, there have been numerous extensions to and deviations from the rational addiction model. Orphanides and Zervos (1995) extend the rational addiction model by incorporating learning about the potential harm associated with consumption of an addictive good. Those authors incorporate heterogeneity in the "addictiveness" of different goods and evolving subjective assessments of this heterogeneity. By consuming a good, agents in the model learn over time the extent to which consumption of that good causes harmful side effects; however, the authors are agnostic as to what these side effects may be. Importantly, the model captures that individuals with an "it won't happen to me" attitude toward consumption of addictive goods are precisely the individuals that may (a.) experience regret and (b.) endure the worst effects of consumption. Alternatively, Suranovic *et al.* (1999) extend the rational addiction model with a specific application to cigarette smoking that makes the withdrawal costs of quitting explicit. In the Becker and Murphy (1988) framework, withdrawal is simply captured

by depreciation in the addictive capital stock. Under positive adjacent complementarity, the marginal utility of addictive consumption is lower if an agent chooses to not consume the good in the previous period. Suranovic *et al.* (1999) impose adjustment costs if the agent reduces consumption but not if she raises consumption. The model better explains why some smokers claim to be "unhappy smokers."

Suranovic *et al.* (1999) is the first paper of which I am aware that explicitly captures the potential health consequences of smoking within a rational addiction framework, albeit in an extremely coarse way . Indeed, in their model, agents are concerned with expected longevity. Agents in the model understand that smoking accumulates in a capital stock that decreases expected longevity at a fixed rate. However, those authors do not allow the smoking stock to depreciate. Thus, the only motivation to quit smoking within their model is to prevent further expected longevity losses. More realistically, Carbone *et al.* (2005) formulate a rational addiction model that is consistent with the health production literature (Grossman, 1972). In their model, agents select from smoking alternatives while internalizing both quality-of-life and death risk effects. Rather than a fixed reduction in expected longevity, death is a stochastic outcome and is a function of both current period and past smoking behavior, as well as an overall health stock. Agents may offset the health effects of smoking with investments in other areas of health. Those authors find that, if agents have access to medical care (i.e. forms of health investment), then agents with a longer life expectancy are more likely to smoke early in life because they anticipate both the depreciation of the health effects of smoking after quitting and offsetting their smoking behavior with health investments later in life.[1]

In addition to numerous extensions, several alternatives to the rational addiction framework have appeared in the literature. As an example of this literature, Bernheim and Rangel (2004) propose a model in which rational economic agents are subjected to environmental cues that may impact their ability to align their choices and preferences. The three main premises of the Bernheim and Rangel (2004) model are: (1.) use among addicts is often a mistake, (2.) experience sensitizes an individual to environmental cues, and (3.) addicts understand and

[1]This result also hinges on time-consistent preferences.

manage their susceptibilities. Given the history an individual brings into a decision making period, as well as some stochastic shock, an agent may depart from the standard rational state or "code mode" and enter "hot mode", in which consumption of an addictive good is compulsory regardless of their contingent plan. Mistakes occur in the model when an individual rationally chooses to abstain from consumption of an addictive good yet, due to an environmental cue (e.g., advertisements, peers pressure, etc.), fails to follow-through with her plan. The environmental cue distribution is a function of the environment choice (e.g., attend a party, stay home, etc.). If the probability distribution of the environmental cues is degenerate such that the environmental cue never exceeds the compulsory use threshold, then the model collapses to a standard rational addiction framework. Other examples of models that deviate from rational addiction include, the time-inconsistent framework of Gruber and Koszegi (2001), in which agents are models as present-biased decision makers, and the nonexpected utility model of Gul and Pesendorfer (2007), in which compulsive consumption is explicitly defined as the variation in consumption in the presence and absence of commitment devices.

Finally, given the nature of the current paper's learning model, the issue of how to define preferences under varying degrees of uncertainty requires further discussion. The von Neumann-Morgenstern axioms that define the expected utility framework dominate the empirical smoking literature. Under this standard framework, agents are indifferent as to the timing of the resolution of uncertainty. Furthermore, a common criticism when empirically testing hypotheses derived from the expected utility framework has been the inverse relationship between the coefficient of risk aversion and the intertemporal elasticity of substitution. Kreps and Porteus (1978) deviate from the expected utility framework by incorporating preferences over the timing of the resolution of uncertainty. A strand of theoretical and empirical literature has flowed from Kreps and Porteus (1978). (See Kreps and Porteus (1978), Kreps and Porteus (1979), and Epstein and Zin (1991)) In the current paper, uncertainty over the individual specific match value is never resolved, rather just lessened (i.e., reduced posterior variance). Implementing nonexpected utility preferences is beyond the scope of this dissertation. Interesting future work might consider several preference specifications (e.g., expected utility, Kreps Porteus preferences, hyperbolic preferences á la Gruber and Koszegi (2001), etc.)

in the context of smoking and addiction.

The large empirical literature stemming from rational addiction has focused on providing evidence of a central tenant of the theory: forward-looking behavior. Indeed, in the rational addiction model, agents must correctly forecast the extent to which consumption of an addictive good today will affect the marginal utility of consumption of that good tomorrow. Initially, much of the empirical literature focused on price effects to provide evidence of forward-looking behavior. Chaloupka (1991) instruments for past and future consumption of cigarettes with past and future prices and state excise taxes and finds a positive relationship between current and lead period smoking behavior. Becker *et al.* (1994) employ a similar identification strategy as Chaloupka (1991) and find, in addition to evidence that both past and future price increases decrease the current demand for cigarettes, that the price elasticity of demand for cigarettes is larger in long-run than the short-run.

More in the spirit of this dissertation, Arcidiacono *et al.* (2007) formulate and estimate various structural models of smoking and drinking behavior that differ on the degree to which individuals are forward-looking. Because cigarette price effects are thought to be small for older individuals (Sloan *et al.*, 2003), Arcidiacono *et al.* (2007) model the tradeoff between current utility from smoking versus the increased probability of future poor health outcomes associated with smoking. Because of this intertemporal tradeoff, how individuals discount the future will shape the "costs" of smoking. For example, a completely myopic individual in this setting will always smoke as long as the instantaneous utility from smoking is greater than from not smoking. In addition to a model estimated for various fixed discount rates, Arcidiacono *et al.* (2007) also estimate the parameters of their model jointly with the discount factor. The discount factor is identified by excluding age from the utility function and exploiting age variation in smoking and heavy drinking in the data. While Arcidiacono *et al.* (2007) are limited by the Health and Retirement Survey to studying the smoking behavior of older individuals, those authors show that forward-looking models better fit their data than do myopic models. Indeed, those authors estimate an annual discount rate of 0.8192 implying a strong degree of forward-looking behavior.

Given the focus on health effects in this dissertation, the next section is devoted to the

literature on how individuals perceive health risks and how information and learning alter those perceptions.

2.2 Information, Uncertainty, and Learning

Extensions of rational addiction that incorporate health effects assume that individuals base their period smoking decisions, in part, on their assessment of the future health effects of smoking. Therefore, one branch of the empirical rational addiction literature has studied the roles of information, risk perceptions, subjective expectations, and learning in the decision to smoke. Are individuals forward-looking with respect to their health and do they correctly forecast risks associated with addictive good consumption?

Viscusi (1990) uses data on individual risk perceptions to assess whether individuals (smokers and non-smokers) have accurate perceptions of the risk of lung cancer.Viscusi (1990) uses data from a telephone survey that asked, among other questions, "Among 100 cigarette smokers, how many of them do you think will get lung cancer because they smoke?" Viscusi compares the response rate from the telephone surveys to risk assessments constructed from the United States Department of Health and Human Services data on the number of lung cancer deaths per year. The results show that both smokers and nonsmokers overestimate the risk of lung cancer associated with smoking. Indeed, if individuals are concerned with avoiding lung cancer, the overstatement of the risk associated with smoking may cause individuals to smoke less if the risk were correctly perceived. Viscusi argues that lung cancer risk perceptions increasing excise taxes on cigarettes has a very similar effect on smoking behavior as increasing lung cancer risk perceptions. These results are confirmed and updated in Viscusi and Hakes (2008).

Of course, assessing the extent to which individuals correctly predict the health consequences of smoking requires that the baseline objective consequences themselves are correct. The researcher cannot assess the accuracy of risk perceptions if the "established" risk itself is biased. In Viscusi (1990), the accuracy of the official risk assessments is not questioned. If,

for example, the effect of smoking on expected longevity is confounded by dynamic selection with respect to smoking, then the loss of longevity estimate will be biased. I will review the literature on expected longevity loss from smoking in the next section.

Furthermore, another criticism of the Viscusi (1990) approach is that the survey questions ask respondents only about the generic smoker and not the individual in question. If individuals have a "it won't happen to me" attitude about health risks, then subjective expectations elicited about the general population risk may be more severe than the individual specific subjective expectations. This distinction is important because when the a rational, forward-looking individual makes the current period consumption decision, she does so based on her subjective assessment of her own risk. Importantly however, these studies do show that individuals are forward-looking and not myopic with respect to the health consequences of smoking decisions. Furthermore, these studies establish that information can change behavior by altering risk perceptions.

Viscusi (1990) also makes a theoretical contribution by modeling an individual's beliefs regarding her health risk from cigarette smoking as a Bayesian function of three factors: a prior risk assessment, some measure of risk from experience (perhaps smoking history, age, etc.), and some new information regarding risk. An important question addressed by the literature has been: what exactly is this new risk information?

One type of new information can be categorized as any information that is directed toward a general audience. A widely publicized example was the landmark 1964 United States Surgeon General report that linked smoking to lung cancer and certain birth defects. Luther L. Terry, then Surgeon General, stated that the report "hit the country like a bombshell. It was front page news and a lead story on every radio and television station in the United States."[2] Did this information deter individuals from taking up smoking? Did smokers at the time respond to the report by quitting? On this question, the literature has been mixed. While much of the literature suggests that informational anti-smoking campaigns decrease cigarette demand for light to moderate smokers, Sloan *et al.* (2003) argue that heavy smokers "do not appear to

[2] http://profiles.nlm.nih.gov/NN/Views/Exhibit/narrative/smoking.html

update these perceptions (on the probability of illness/death due to smoking) in response to general information; they need the message to be *personalized*."[3]

Personalized health information may be an important motivator to quit if heavy smokers possess an "it won't happen to me" attitude. Khwaja *et al.* (2006), studying individuals from the Health and Retirement Survey (HRS), show that smokers "learn" about the risks associated with smoking, as measured by a change in smoking behavior, only from a shock to their own health. Wray *et al.* (1998) find a similar result and note that, following a heart attack, there is a strong positive relationship between the propensity to quit and education. Khwaja *et al.* (2006) argue that if any health shock other than one's own would encourage smoking cessation, it should be that of a spouse. The authors however find no significant effect of spousal health shocks on smoking behavior. Sloan *et al.* (2003) conclude that "the clear differences in the effects of smoking-related health shocks for current smokers suggest that personalized messages, relevant to their circumstances, are necessary to get their attention and induce changes in their beliefs (qtd. in Sloan *et al.* (2003) pg. 124)."

Nearly all previous work that has examined learning or expectation formation with respect to personalized health messages has studied behavioral changes after a *major* health shock to self or spouse (Smith *et al.*, 2001; Khwaja *et al.*, 2006; Arcidiacono *et al.*, 2007). Additionally, most papers focus on individuals above the age of 50, at which age we begin to observe the major health implications of smoking. For example, using Health and Retirement Survey data, Smith *et al.* (2001) show that current smokers update their subjective longevity expectations given a chronic, *smoking-related* health shock (e.g., certain cancers, cardiovascular disease, etc.) differently than do former and non-smokers. Those authors find no difference in longevity updating for health shocks that are unrelated to smoking (e.g., prostate cancer). Their study concludes, "it remains to be evaluated whether messages can be designed that focus on the link between smoking and health outcomes in ways that will have comparable effects (to chronic health shocks) on smokers' risk perceptions."

[3]Italics theirs.

This dissertation explores the notion that personalized health information does not necessarily have to come in the form of a major health shock after age 50. Indeed, as recognized by the literature, waiting for a major health shock to incite individuals to quit smoking may be too late in terms of life expectancy gains. Therefore, one goal of this dissertation is to examine the extent to which personalized health marker measures (e.g., blood pressure, cholesterol, etc.) at all ages might inform individuals of the health risks associated with smoking. As shown above, the medical literature suggests that a number of health markers that predict many chronic health outcomes may be influenced by smoking cigarettes. Health markers are theorized to provide tangible evidence to both the "it won't happen to me" and the light smoker of the health risks associated with smoking without being a one-off, potentially fatal, signal of information. For example, the Surgeon General now recommends that physicians point out to patients the fact that smoking raises blood pressure.[4] To my knowledge, no study has examined the impact of personalized health marker information on the decision to smoke. The extent to which information on these health markers may affect the decision to smoke are central questions of this research .

2.3 Smoking, Health, and Selection

Behind much of the United States Surgeon General Report's results is the work of (Doll *et al.*, 1994, 2004). Those authors use survey data of British physicians over several decades to assess the impact of cigarette smoking on mortality. For different birth cohorts, those authors construct the relative risk of death, defined as the age specific mortality for a group in question relative to some baseline group. Their findings suggest that smoking cessation at ages 30, 40, 50, and 60 leads to improved life expectancies of 10, 9, 6, and 3 years respectively. Furthermore, life-long smokers face a roughly 25 percentage point increase in the probability of death during middle aged (35-69). They conclude that, for men born between 1900 and 1930 who smoked only cigarettes, the mean reduction in longevity was ten years relative to a group of life long nonsmokers. However, those authors do not consider the intensity with which one

[4] http://www.cdc.gov/tobacco/data_statistics/sgr/2010/clinician_sheet/pdfs/clinician.pdf

smokes in any of their reported results. As I show in this dissertation, the intensity with which one smokes greatly influences the potential loss of longevity.

Taylor *et al.* (2002) use data from the Cancer Prevention Study II to assess the life-years lost from smoking and the longevity benefits of smoking cessation at various ages. They estimate the relative risk of mortality for different categories of smokers. Unfortunately, after an initial interview in 1982, longitudinal data is only available for a self-selected fraction of the baseline sample. After attempting to correct for misclassification of mortality due to changes in smoking behavior over the study period, the authors find that men that never smoke have an mean expected longevity that is 8.9 to 10.5 years longer than those that smoke until death. For women the life extension is 7.4 to 8.9 years. Furthermore, Taylor *et al.* (2002) find that there are expected longevity gains from smoking cessation at all ages. For example, individuals that smoke until age 65 and then quit, gain about 1 and 2.7 year(s) of life expectancy for men and women respectively.

Employing a similar relative risk methodology to Doll *et al.* (2004), Brønnum-Hansen *et al.* (2007), using mortality data from Danish Health Interview Survey find that expected longevity from age 25 is reduced by 8.7 years for men and 10.4 years for women for heavy smokers ($\geq$ 15 cigarettes/day) relative to life long nonsmokers. Additionally, those authors find that heavy smokers had 10.5 fewer years without chronic "limiting" illnesses relative to life long nonsmokers. Taking into account secondary factors, Jha *et al.* (2006) focus on smoking attributed mortality differences by overall socioeconomic status. They find that more than half of the absolute difference in mortality rates can be attributed to smoking-related health problems and that smoking contributed to over 40% of middle-aged deaths (35-69) for the lowest socioeconomic group.

In assessing the effect of smoking cigarettes on expected longevity, an important consideration is the extent to which selection and endogeneity play roles. For example, each of the above studies ignores the potential correlation between preferences for smoking cigarettes and the mortality outcomes in question. Perhaps, individuals that are more likely to enjoy smoking (e.g., because of parental effects, stress, environmental cues, etc.) are also more likely to die for other reasons independently of smoking (e.g., alcohol and other drug consumption, risky

lifestyle behaviors, etc.). In effect, these studies are treating smoking as a random event, in direct odds with the rational addiction framework above. Furthermore, these studies ignore the possibility of health selection in smoking behavior. If, independently of smoking, smokers are of a worse overall health status than non-smokers, standard statistical methods may overstate the effect of smoking on mortality. Indeed, only recently have papers in health economics begun to jointly model smoking and health outcomes. For example, Adda and Lechene (2001) show that potential life-span and smoking behavior are correlated along unobserved (to the econometrician) dimensions. However, Adda and Lechene (2001) estimate survival probabilities independently of smoking and fix key depreciation parameters. Adda and Lechene (2004) try to control for the selection problem by constructing a health index that proxies for the health of a smoker had the individual not smoked. Those authors are not able test whether cigarette smoking changes in behavior over time influence their health index because they have a very short panel.

2.4 Empirical Methods

One contribution of this dissertation is the empirical treatment of the theoretically common "smoking stock." In the rational addiction framework, an individual's smoking stock is a continuous, scalar representation of all past smoking behavior. Becker and Murphy (1988) specify this history as a capital stock that is subject to depreciation and withdrawal. To test a central tenant of the rational addiction theory, the stock must be constructed such that it captures how all past consumption (smoking) influences the marginal utility of current consumption.[5] One candidate for the smoking stock state variable is simply the number of years smoked in the past. For example, consider a longitudinal dataset that includes whether an individual smoked and the intensity with which he/she smoked in each period. A simple stock specification could simply be the total number of past periods in which an individual smoked. However, this

[5]For the purposes of this discussion, I will use smoking as the running example. As noted in Becker and Murphy (1988), a wide variety of goods can be considered addictive and this discussion applies to these goods as well.

specification ignores the possibility that the timing of when an individual smoked is important in predicting current smoking through the marginal utility. That is, two individuals may have each smoked for 10 periods, but one individual smoked more recently than the other. Furthermore, this specification ignores important information with respect to the intensity with which someone smokes (i.e., one cigarette per day versus one pack of cigarettes per day).

To capture both duration and timing, as well as intensity of smoking, I use a data reduction method, principal component analysis (PCA), to construct the stock variable. Clearly, in the context of structural models, reducing the dimension of the state space is advantageous. However, Sickles and Williams (2008) is the only paper of which I am aware that uses a data reduction method to construct a state variable in a dynamic structural model. In their application, overall individual capital is decomposed into a standard human capital stock and a "social" capital stock. Both capital stocks are then used to help explain observed patterns of criminal behavior at both the extensive and intensive margins. The social capital stock is constructed from past criminal behavior, labor force variables, and demographic characteristics to account for the social norms that may influence crime related decisions. Importantly, and similarly to this dissertation, those authors are able to use data to construct a theoretically well defined term: social capital.

Before proceeding, I present a overview of principal component analysis and a brief example. Consider a dataset with k variables and n observations. In matrix notation, denote $X_{n\times k}$ as these data. The total variation in the data can be summarized via the product $X'X$, which is a square $k \times k$ matrix. Given the k dimensions and the fact that the matrix is square, we can (normally) derive k eigenvectors of the matrix. Each eigenvector explains an orthogonal part of the total variation in $X'X$. Associated with each eigenvector is an eigenvalue that scales the eigenvector. Because the sum of the k eigenvalues equals k, each eigenvalue represents the proportion of variance explained by its corresponding eigenvector. Sorting the eigenvectors by the corresponding eigenvalues from largest to smallest ranks the eigenvectors by the amount of total variation explained in $X'X$. The key attribute is that the eigenvectors are orthogonal. Thus, each eigenvector explains a different part of the total variation in $X'X$. The first principal component is constructed as a linear combination of the k variables and

the associated factor loadings from the eigenvector with the highest eigenvalue. PCA is a non-parametric method that restricts the squared eigenvector components to sum to one. As an example, suppose a researcher wants to construct a wealth index. The available data have 40 variables that are loosely related to wealth. The researcher may employ PCA to reduce the dimension of this information. Suppose the first 3 principal components explain 95% of the total variation in the 40 variables. The researcher could use the first three PCs in his analysis without losing much information and while greatly simplifying the computational burden of the analysis. Of course, choosing the number of principal components to use is more art than science. I postpone the discussion of my specific smoking stock treatment until Chapter 5.

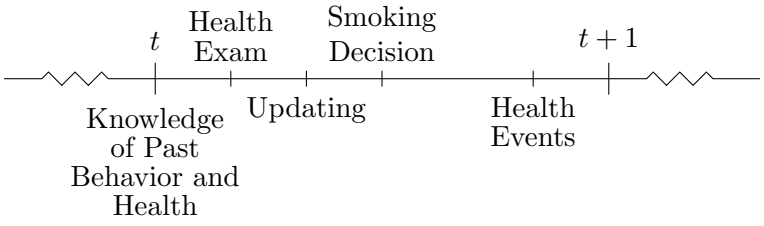
Finally, in this dissertation, estimation of the primitive parameters of one's decision making optimization problem (e.g., preferences, constraints, and expectation parameters) allows me to assess the impact of smoking on morbidity and mortality outcomes while considering the potential for endogeneity of and dynamic selection into smoking behaviors. In addition, the introduction of serially correlated (permanent) unobserved heterogeneity that affects decision making over the life cycle (including observed initial conditions) allows for the recovery of parameters that measure the impact of smoking and health markers on morbidity and mortality that are free of selection bias. Unbiased estimation of these primitive parameters allows me to simulate the model and impose different patterns of smoking and quitting to examine the resulting changes in predicted health outcomes. Finally, my model builds on the Bayesian learning structural models of other papers in pharmaceutical demand (Crawford and Shum, 2005; Chan and Hamilton, 2006), fertility and infant mortality (Mira, 2007), marketing (Ackerberg, 2003), and health plan report cards (Chernew *et al.*, 2008). I postpone a thorough review of my solution and estimation methods until Chapters 3 and 5, respectively.

Chapter 3

Theoretical Model

This chapter describes both the main theoretical model of the dissertation and the solution methods used to solve the model. When necessary, I discuss the econometric structure that facilitates estimation within this section; however, I postpone the comprehensive discussion of my specific estimator, simulated maximum likelihood, to chapter 5. Furthermore, changes to the model that are required due to data limitations are discussed in chapter 5.

I specify a dynamic stochastic model of smoking behavior that incorporates learning. Consider a mixed discrete/continuous-state, discrete-time model in which a the time subscript t corresponds to individual i's age . The relevant time frame is from age seven until death. The model has a finite horizon in the sense that, while I model an explicit stochastic probability that an individual may die prior to period T, the probability of death equals one in period T.[1] Let a representative period t follow the following time line.



[1] T is set to age 100. Finite horizon models are common in the dynamic discrete choice structural model literature because iterating on the value function is computationally burdensome. See (Rust and Phelan, 1997; Aguirregabiria and Mira, 2010) for a discussion.

An individual enter the period with knowledge of her past smoking behavior and their past health. Next, an individual undergoes a health exam in which she learns some personalized information about a set of her health markers. Using the health marker information, an individual then updates her beliefs on how her smoking behavior has affected her health markers. Conditional on her updated beliefs and other state variables, the individual then makes her period t smoking decision. Finally, if any chronic health or mortality events occur, they are assumed to take place at the end of the period. In the following sections, I describe each of these points in greater detail.

3.1 Knowledge Entering the Period

Following the rational addiction literature, A_{it} represents the accumulated smoking "stock". The concept of a smoking stock is not immediately intuitive. Broadly speaking, the rational addiction literature treats the stock as a measure of past smoking. Medically, however, we might consider the stock as some accumulation of tar in the lungs that influences health. Alternatively, we might think of the stock as a measure of dependence on nicotine. These alternative definitions are suppressed in the current model in favor of a stock variable that may be interpreted as a continuous summary of an individual's smoking history. Here, the extent to which A_{it} influences health and future smoking probabilities is dependent on the model parameters, and thus an empirical question. Formally, the stock is defined as:

$$A_{it} = \left\{ \begin{array}{ll} \exp\left\{\delta_1 \ln(A_{it-1}) + \delta_2 \mathbf{1}[d_{it-1}=1] + \delta_3 \mathbf{1}[d_{it-1}=2] + \rho^A \mu_i + \eta_{it}\right\} & \text{if} \quad \sum_{n=1}^{t-1} d_{in} > 0 \\ 0 & \text{otherwise} \end{array} \right\} \tag{3.1}$$

Equation 1 says that individual i's time t smoking stock is normalized to 0 if she has not smoked in *any* previous period. Conditional on any past smoking, the stock is specified as a function of the previous period stock and the previous period decision, to be defined below. δ_1 can be interpreted as one minus the depreciation rate of the stock in percentage terms. The nonlinear investment of light and heavy smoking into the smoking stock are captured by δ_2 and δ_3, respectively. Unobserved and individual specific permanent heterogeneity is

captured by the μ_i term and its factor loading ρ^A.[2] Also influencing the stock is an i.i.d. white noise term, η_{it}, which is distributed $\mathcal{N}(0, \sigma_\eta)$.[3] Consistent with the interpretation of the stock as a summary of an individual's smoking history, the stock is assumed to be known by the individual in each period.

Also known to the individual at the beginning of period t is her overall chronic health state, H_{it}. Let $H_{it} = h$, where outcome h is as follows:

$$h = \left\{ \begin{array}{lll} 1 & \text{if} & \text{Chronic Condition} \\ 0 & \text{if} & \text{No Chronic Condition} \end{array} \right\}$$

While I postpone a more complete definition of possible chronic conditions to chapter 4, the chronic health state can be thought of including all types of cancers and forms of cardiovascular disease. For the purposes of the theoretical model, a chronic condition captures some of the potential consequences of smoking. As I show below, smoking influences the probability of contracting a chronic condition which, in addition to smoking, affects the probability of death. Furthermore, as will be seen in the utility function, preferences for smoking may vary depending upon an individual's chronic health state. While aggregating the data into a dichotomous indicator for chronic is a coarse treatment of health, this method allows for chronic health to affect other probabilities and preferences within the model in a computationally tractable way. Furthermore, the chronic health state need not be interpreted as the set of diseases caused by smoking. Rather, individuals in the model understand that smoking cigarettes simply raises the probability of transiting to the chronic health state. This interpretation is consistent with the treatment of health in Arcidiacono *et al.* (2007), where "bad health" is simply the absence of self-reported "good health".

---

[2] See section 5 for a discussion of estimation and interpretation issues regarding the permanent unobserved heterogeneity.

[3] Given the exponential stock evolution equation, η is a log normal shock.

3.2 The Health Exam

Next, an individual attends a health exam and receives personalized health information about health markers (e.g., blood pressure, cholesterol, etc.). Define R_{it} as a continuous scalar summary of these health markers. R_{it} is assumed to evolve as follows:

$$R_{it} = \zeta R_{it-1} + X_{it}\phi + \kappa_{it} + \rho^R \mu_i. \tag{3.2}$$

Here, X_{it} is a vector of sociodemographic characteristics of individual i. I assume that the technology associated with these characteristics (i.e., ϕ) is known by the individual. ζ captures the dynamic aspect of the health markers and is also assumed to be known to the individual. Time invariant and unobserved (to the econometrician) heterogeneity is captured by the μ_i term and its factor loading ρ^R. Let κ_{it} represent the input from the smoking stock plus an idiosyncratic, i.i.d. error term that is defined as:

$$\kappa_{it} = \theta_i A_{it} + \nu_{it}. \tag{3.3}$$

Because the individual observes or knows R_{it}, ζ, $X_{it}\phi$, and $\rho^R\mu_i$, κ_{it} is also observed by the individual. I allow there to be heterogeneity in the effect of the accumulated smoking stock, A_{it}, on the summary measure of health markers, R_{it}. I theorize that each individual is endowed with a time invariant, unknown (to both the individual and econometrician) match value θ_i that captures this idiosyncratic relationship. θ_i is drawn from a known population distribution given by:

$$\theta_i \sim \mathcal{N}(\overline{\theta}, \sigma^2_\theta).$$

κ_{it} therefore serves as a noisy signal of information. Over time, by having health exams and thus observing a sequence of signals, θ_i is learned in a Bayesian fashion. Learning is, however, confounded by the i.i.d. noise term, ν_{it}. Indeed, without ν_{it}, an individual would perfectly learn their match value θ_i at the first health exam (i.e., the first realization of κ_{it}). While ν_{it}

is unknown, its distribution is known and given by:

$$\nu_{it} \sim \mathcal{N}(0, \sigma_\nu^2).$$

Because θ_i is time invariant, and because the distributions of θ_i and ν_{it}, as well as the stock A_{it}, are known, over time, an individual can learn her idiosyncratic value of θ_i.[4]

3.3 Learning

Let an individual's period t posterior beliefs, those with which she forecasts future health markers, be given by τ_{it}, her posterior mean, and ψ_{it}, her posterior variance. I assume rational expectations such that an individual's initial belief, prior to any health exams, regarding her true θ_i (the marginal effect of one's smoking history, A_{it}, on health markers, R_{it}) is the population distribution.[5] Initial beliefs ($t = 0$) are:

$$\tau_{i0} = E_0(\theta_i) = \overline{\theta}$$

$$\psi_{i0} = V_0(\theta_i) = \sigma_\theta^2.$$

Expectations about future health marker transitions evolve in the current model with the receipt of personalized health information. In deriving posterior beliefs, consider an individual in period t with smoking stock A_{it}. This individual has two fundamental sources of information: her prior beliefs, $(\tau_{it-1}, \psi_{it-1})$, and the observed results from her period t health exam, κ_{it}. Appealing to the assumption of conjugate prior and signal distributions, the period t beliefs have closed form solutions that are given via Bayes' Rule. The posterior mean and variance

[4]The assumption that an individual knows the technology of the health production function is ubiquitous in health economics. That is, typically $\theta_i = \overline{\theta} \; \forall \; i$ and $\sigma_\theta = 0$. $\overline{\theta}$ is then estimated and assumed to be the marginal product that all individuals use to solve optimization problems.

[5]The rational expectations assumption is what is typically made in most models of health transitions.

are :[6]

$$\tau_{it} = E(\theta_i|\kappa_{it}, A_{it}, \tau_{it-1}, \psi_{it-1}) = \frac{A_{it}^2\psi_{it}}{\sigma_\nu^2}\hat{\theta}_{it} + \frac{\psi_{it}}{\psi_{it-1}}\tau_{it-1} \tag{3.4}$$

$$\psi_{it} = Var(\theta_i|A_{it}, \psi_{it-1}, \sigma_\nu) = \frac{\psi_{it-1}\sigma_\nu^2}{A_{it}^2\psi_{it-1} + \sigma_\nu^2}. \tag{3.5}$$

Here, $\hat{\theta}_{it}$ is the least squares estimate of κ_{it} on A_{it} from the within individual variation of the t^{th} health exam. Note that these beliefs have the following appealing properties. First, the posterior mean is a weighted average of $\hat{\theta}_{it}$ and the original prior mean τ_{it-1}. Second, the weight placed on the period t signal (i.e., $\hat{\theta}_{it}$) is increasing in the smoking stock. Finally, the posterior moments of an individual for whom the stock equals zero (i.e., $A_{it} = 0$) collapse to the prior moments.

3.4 The Smoking Decision

Each period, a forward-looking individual makes a smoking decision to maximize her lifetime discounted expected utility. Let the decision for individual i be given by $d_{it} = d$, where smoking alternative d is:[7]

$$d = \left\{ \begin{array}{ll} 0 & \text{Do not smoke} \\ 1 & \text{Smoke} \leq 1 \text{ Pack/day} \\ 2 & \text{Smoke} > 1 \text{ Pack/day} \end{array} \right\}$$

The set of factors that influence individual $i's$ smoking decision in period t are given by the state space S_{it}. Define S_{it} as follows:

$$S_{it} = \{A_{it}, R_{it}, \tau_{it}, \psi_{it}, H_{it}, X_{it}\}$$

where A_{it} is individual $i's$ smoking stock entering period t; R_{it} is her health marker index; τ_{it} and ψ_{it} are her mean and variance respectively of her posterior belief distribution; H_{it} is her

[6] Derivations of these equations can be found in Appendix A.

[7] Aggregating the smoking decision to this level is a simplifying assumption. However, the model is still able the important distinction between light and heavy smoking that is often ignored in the medical literature.

chronic health status; and X_{it} is her set of demographic characteristics. Additionally influencing behavior, but not listed here, are a preference error ϵ_{it} and a permanent heterogeneity term μ_i that are both assumed to be known to the individual but *unobserved to the econometrician.*

Following the standard expected utility framework, an individual makes the smoking decision by evaluating the lifetime present discounted expected value of each smoking alternative and selecting the highest valued alternative. The deterministic portion of per-period utility associated with health state h, $(h = 0, 1)$ and smoking alternative $d_{it} = d$ is:

$$
\begin{aligned}
&\overline{U}^h(S_{it}, d_{it} = d, \mu_i) = \\
&\quad \alpha_{0h} \\
&\quad +(\alpha_{1h} + \alpha_{2h}A_{it} + \alpha_{3h}R_{it} + \alpha_{4h}Age_{it}) * \mathbf{1}[d_{it} = 1] \\
&\quad +(\alpha_{5h} + \alpha_{6h}A_{it} + \alpha_{7h}R_{it} + \alpha_{8h}Age_{it}) * \mathbf{1}[d_{it} = 2] \\
&\quad +\alpha_{9h} * \mathbf{1}[d_{it-1} \neq 0] * \mathbf{1}[d_{it} = 0] \\
&\quad +\alpha_{10h}A_{it} + \alpha_{11h}A_{it}^2 \\
&\quad +\rho^{Uhd}\mu_i
\end{aligned}
\tag{3.6}
$$

First, note that all utility parameters are chronic health state specific and the value of death is normalized to zero. Because there is no explicit outside consumption good, the utility framework focuses on how smoking and health shift utility from some baseline levels, $\alpha_{0\cdot}$. The specification accommodates any nonlinearity in the effects of light and heavy smoking on utility. While $\alpha_{1\cdot}$ ($\alpha_{5\cdot}$) is the direct marginal utility of light (heavy) smoking, $\alpha_{2\cdot}$ ($\alpha_{6\cdot}$) captures the extent to which past consumption reinforces current consumption. $\alpha_{2\cdot}$ ($\alpha_{6\cdot}$) captures a part of the intertemporal trade-off in utilities. The extent to which the health marker index affects the marginal utility of smoking is captured by $\alpha_{3\cdot}$ ($\alpha_{7\cdot}$). Note that higher values of R_{it} and

A_{it} imply worse health and a higher smoking stock respectively. The signs and magnitudes of $\alpha_{2\cdot}$ $\alpha_{3\cdot}$, $\alpha_{6\cdot}$, and $\alpha_{7\cdot}$ are empirical questions. $\alpha_{4\cdot}$ ($\alpha_{8\cdot}$) captures changes in the marginal utility of smoking across the lifespan. Specific withdrawal costs from quitting, which also capture part of the intertemporal utility trade-off, are captured by $\alpha_{9\cdot}$. Finally, $\alpha_{10\cdot}$ and $\alpha_{11\cdot}$ capture tolerance in smoking. That is, the extent to which a given level of stock affects utility is captured here regardless of smoking behavior.

Relative preferences over smoking alternatives hinge on two main factors. First, preferences vary by the chronic health state ($H_{it} = h$). The extent to which the marginal utility of smoking varies across chronic health states remains an open question. Generally, the marginal utility of consumption of any normal good is thought to be lower in worse health states (Viscusi and Evans, 1990; Gilleskie, 1998). If however smoking provides relaxation and comfort when stricken with a chronic illness, the overall marginal utility of smoking may be larger in worse health states. Estimation of the structural parameters will therefore empirically test for the sign of the marginal utility of smoking across health states. Second, as seen in equations for the smoking stock and health marker index, current period smoking affects the size of the next period smoking stock, which in turn affects the next period health marker index and next period utility. Given the dynamic nature of the model, individuals evaluate smoking alternatives while considering the future marginal utility of smoking as well as the future consequences of a higher A_{it}.

Following Rust (1987), let the total current period utility be the sum of the deterministic utility from equation 3.6 and an additive i.i.d. preference shock that is alternative and health-state specific:

$$U^h(S_{it}, d_{it} = d, \mu_i, \epsilon_{it}^{dh}) = \overline{U}^h(S_{it}, d_{it} = d, \mu_i) + \epsilon_{it}^{dh}.$$

In the empirical implementation below, ϵ_{it}^{dh} is simply an additive econometric error; however, in the theoretical model, ϵ_{it}^{dh} is given a structural interpretation as an unobserved state variable (Aguirregabiria and Mira, 2010). The alternative specific lifetime value function, conditional

on unobserved heterogeneity μ_i, is:

$$V_d^h(S_{it}, \epsilon_{it}^{dh}|\mu_i) = \overline{U}^h(S_{it}, d_{it} = d, \mu_i) + \epsilon_{it}^{dh} + \beta\big[(1 - \varsigma_{t+1}) \sum_{a=0}^{1} \pi_{t+1}^a E_t[V^a(S_{it+1}|\mu_i)|d_{it} = d]\big]. \tag{3.7}$$

Here, the value function, $V^a(S_{it+1}|\mu_i)$, is the maximal lifetime utility in period $t+1$ of being in chronic health state a. The explicit probabilities ς_{t+1} and π_{t+1}^a that alter the future value function are the probabilities of death and chronic health state a respectfully and are defined below. The value function is conditional on the unobserved heterogeneity component μ_i. The expectation operator is taken over the distribution of the individual match value θ_i, the smoking stock and health marker state variables, *and* the preference error ϵ. To evaluate this integral, I proceed in two steps.[8] First, let $\overline{V}^h(S_{it}|\mu_i) = \int V^h(S_{it}|\mu_i) dG_\epsilon(\epsilon_{it}^{dh})$, where $G_\epsilon(\epsilon_{it}^{dh})$ is the cumulative distribution function of ϵ. By assuming that ϵ_{it}^{dh} is i.i.d. Extreme Value Type I distributed, the maximal (EMAX function) expected lifetime utility has the following closed form solution:

$$\overline{V}^h(S_{it+1}|\mu_i) = EC + ln\Big(\textstyle\sum_{d=1}^{D} exp\big(\overline{V}_d^h(S_{it+1}|\mu_i)\big)\Big) \quad \forall t, \quad \forall h. \tag{3.8}$$

Here, EC is Euler's constant. However, I still must integrate over future state transitions and beliefs. Thus, in the second step, given the unitary dimension of the posterior distribution, as well as the i.i.d. nature of other shocks in the model (smoking stock and health marker error terms), I use a Monte Carlo method to evaluate the expectation within solution to the model (described below). Finally, if we assume the error term ϵ_t^d is additively separable, then we can decompose the choice-specific value function to be $\overline{V}_d^h(S_{it}|\mu_i) = V_d^h(S_{it}, \epsilon_{it}^{dh}|\mu_i) - \epsilon_{it}^{dh}$ and the conditional choice probabilities take the following dynamic multinomial logit form:

$$p(d_{it} = d|S_{it}, \mu_i) = \frac{exp\big(\overline{V}_d^h(S_{it}|\mu_i)\big)}{\sum_{d=0}^{2} exp\big(\overline{V}_d^h(S_{it}|\mu_i)\big)} \ \forall t, h \tag{3.9}$$

[8] This procedure is possible because the preference error ϵ is i.i.d. across choices, health states, and time and enters the utility function additively.

To preview the empirical implementation, the conditional smoking choice probability in equation 3.9 enters the likelihood function. The parameters that dictate the choice probability are structural in the sense that they are follow from the above maximization problem.

3.5 Chronic Health and Mortality

At the end of the period, with some probability an individual may transit to the chronic health state. What differentiates H_{it} and R_{it} is "reversibility". While R_{it} changes each period, I assume that upon diagnosis of a chronic condition, an individual has the condition forever.[9] Let the probabilities of transiting to different chronic health states in period $t+1$ be:

$$\pi^0_{t+1} = \left\{ \begin{array}{ll} [1 - P(H_{t+1} = 1|S_{it}, d_{it}, \mu_i)] & \text{if } H_{it} = 0 \\ 0 & \text{if } H_{it} = 1 \end{array} \right\}$$

$$\pi^1_{t+1} = \left\{ \begin{array}{ll} P(H_{t+1} = 1|S_{it}, d_{it}, \mu_i) & \text{if } H_{it} = 0 \\ 1 & \text{if } H_{it} = 1 \end{array} \right\}.$$

Define the relevant probability, $P(H_{t+1} = 1|S_{it}, d_{it}, \mu_i)$, with the following binary logit equation:

$$\frac{exp(\lambda_0 + \lambda_1 R_{it} + \lambda_2 R^2_{it} + \lambda_3 \mathbf{1}[1980s] * R_{it} + \lambda_4 \mathbf{1}[1990s] * R_{it} + [\lambda_5 + \lambda_6 R_{it}] * d_{it} + \lambda_7 X_{it} + \rho^H \mu_i)}{1 + exp(\lambda_0 + \lambda_1 R_{it} + \lambda_2 R^2_{it} + \lambda_3 \mathbf{1}[1980s] * R_{it} + \lambda_4 \mathbf{1}[1990s] * R_{it} + [\lambda_5 + \lambda_6 R_{it}] * d_{it} + \lambda_7 X_{it} + \rho^H \mu_i)} \tag{3.10}$$

Here, R_{it} is the health marker index defined above, X_{it} is a vector of exogenous individual characteristics, d_{it} is the smoking choice and μ_i is an individual, time invariant unobserved heterogeneity term. The factor loading superscript H simply differentiates it from other factor loadings in the model. λ_6 and λ_7 capture changes over time in how health markers affect the probability of chronic disease incidence (perhaps due to advances in medical technology, pharmaceuticals, etc.).

In forecasting future chronic health transitions, I follow the literature and assume that an

[9] This assumption captures the fact that upon having an heart attack, for example, an individual is in a fundamentally different health state even if they don't have repeated heart attacks (Khwaja *et al.*, 2006).

individual has rational expectations and that she understands the technology associated with the chronic health transition probability. A natural question becomes, why do individuals in the model learn about how smoking affects health markers but not chronic conditions? By modeling learning about the effect of A_{it} on R_{it}, however, individuals are indirectly updating their expectations about future chronic health transitions because the health marker index enters the chronic health transition probability. Furthermore, the purpose of this paper is to explore the importance of health information prior to major health shocks. Imposing that individuals understand the technology (i.e., the λs) associated with covariates in the chronic health transition equation is the standard approach. While future work may incorporate learning about health transition probabilities, such learning is currently beyond the scope of this paper.

Finally, conditional upon her period t smoking decision, her realized chronic health state, H_{it+1}, and other state variables, S_{it}, an individual may die at the end of period t. Define an indicator for death at the end of period t, M_{it+1}=1, and let its corresponding probability, $\varsigma_{t+1} = P(M_{it+1} = 1|S_{it}, d_{it}, \mu_i)$, be given by:

$$\frac{exp(\omega_0+\omega_1 R_{it}+\omega_2 R_{it}^2+\omega_3 H_{it+1}+[\omega_4+\omega_5 R_{it}+\omega_6 H_{it+1}]*d_{it}+\omega_7 \mathbf{1}[1980s]*H_{it+1}+\omega_8 \mathbf{1}[1990s]*H_{it+1}+\omega_9 X_{it}+\rho^M \mu_i)}{1+exp(\omega_0+\omega_1 R_{it}+\omega_2 R_{it}^2+\omega_3 H_{it+1}+[\omega_4+\omega_5 R_{it}+\omega_6 H_{it+1}]*d_{it}+\omega_7 \mathbf{1}[1980s]*H_{it+1}+\omega_8 \mathbf{1}[1990s]*H_{it+1}+\omega_9 X_{it}+\rho^M \mu_i)} \tag{3.11}$$

Here, H_{it+1}, is individual $i's$ chronic health state at the end of period t.[10] Again, the superscript on the factor loading simply differentiates it from other factor loadings. The technology for the death transition equation is assumed to be known by the individual. ω_7 and ω_8 capture the fact that, conditional on having some chronic illness, the probability of death from that illness may have changed over time due to medical advances. Furthering the discussion above, because the health marker index enters the death transition equation directly (and indirectly through the chronic health term H_{it}), individuals are indirectly updating their expectations

[10]The timing convention here is due to data aggregation. Clearly, any chronic health event occurring in period t must occur at or before the time of death, if death also occurs in t. Therefore, to accommodate the frequent observation in the data of an individual dying from a chronic health event, the appropriate chronic health data point in this equation is H_{it+1}.

about death transitions conditional on their smoking choice through the learning process. Assuming that the ωs are known by the individuals is the standard approach and one that can be relaxed in future work.

After the realization of the chronic health and mortality outcomes, conditional on remaining alive, time transitions to period $t+1$. The decision making process is consistent in the sense that, conditional on an individual's state, the problem to be solved is the same across time. Furthermore, the fixed discount factor rules out any time inconsistency with respect to preferences. Solution of the model amounts to characterizing the integrated Bellman equation (EMAX) in equation 3.8. I will discuss the solution and estimation method together in Chapter 5.

Chapter 4

Description of The Framingham Heart Study

The Framingham Heart Study is one of the longest running panel studies in the world. With the stated goal to "identify the common factors that contribute to cardiovascular disease", the Study contains repeated observations of individuals over a 50 year period.[1] Research milestones from the Study range from the causal relationship between cigarette smoking and heart disease in the early 1960s to the effect of obesity on heart failure and the lifetime risks of contract various cardiovascular ailments. Indeed, the Study has resulted in over 1200 scholarly articles being published in medical journals.

Beginning in 1948, the Framingham Heart Study began collecting biennial health data from 5,209 individuals living in Framingham, Massachusetts. These individuals formed what became known as the Original Cohort. Participants underwent physical examinations as well as lifestyle interviews. While the nature of these examinations has changed dramatically over the study period, there has been a constant focus on cardiovascular health. In 1971, the Framingham Heart Study began following the offspring and spouses of the Original Cohort to form the Offspring Cohort. Each cohort, the Original and the Offspring, represents a different panel study that has continued into the 21st century. While lacking income information and geographical variation, the data contain a wealth of health and smoking information that

[1] The Framingham Heart Survey: http://www.framinghamheartstudy.org/index.html

are ideal for analyzing the tradeoff between smoking and the potential for future health shocks.

4.1 Sample Construction

The structural model in chapter 3 is estimated with data from the Framingham Offspring Cohort.[2] The decision to focus on data from the Offspring Cohort stems from the consistency with which the health exams were administered. Smoking and health questions changed substantially over time in the Original Cohort; thus, constructing uniform measures of smoking history, per-period behavior, and health variables (especially health markers) proved to be difficult. In constructing the sample used in estimation, I drop all individuals with a missing exam and all those lost to attrition. Those with missing exams were dropped because, given the length of time between exams, assessing smoking behavior during the interim period proved to be difficult. An additional 671 individuals are lost to attrition (i.e., some reason other than death) at some point during the seven exams. Those that attrit from the sample constitute approximately 18% of the sample. The decision to drop these individuals is based on the computational tractability of modeling attrition. Simple t-tests for difference of means suggest that those that attrit are slightly more likely to be women, have a three point lower level of systolic blood pressure on average, and have a statistically insignificant difference in coronary heart disease incidence than their nonattriting counterparts. Those that attrit are on average slightly more likely to smoke. Thus, due to the restrictions I have placed on the sample, the *only* reason individuals leave the sample is through death. As shown in Chapter 3, the probability of death is explicitly modeled.

Table 4.1 explains my process of sample construction. The full sample contains 5,124 individuals. For this work, I only have access to data for those individuals from whom consent for distribution was granted. The final sample consists of 19,461 person/year observations.

[2]In another study, I am examining the intergenerational transfer of smoking preferences between Original and Offspring Cohort participants.

Table 4.1: Sample Construction

N	Description
4989	Framingham Heart Survey Offspring Cohort Participants - Restricted Sample
3730	Sample after dropping those individuals that skipped one or more of the health exams
3008	Sample after dropping all person/year observations of individuals who attrit

3008 unique individuals yields 19461 person/year observations.

Source: The Framingham Heart Study, Offspring Cohort.

4.2 Sample Statistics

The sample statistics given in this section are by Framingham Heart Study exam. I postpone the reconciliation of the data and the theoretical model to chapter 5. I have data for each individual in the Framingham Heart Survey Offspring Cohort for up to seven health exams. Initial health exams were conducted between 1971 and 1975. For each participant, subsequent exams occurred at varying time intervals. Table 4.2 provides information on the average timing of each exam across individuals, in addition to demographic information. Because attrition

Table 4.2: Sample Characteristics by Exam

Exam	Mean Year	Mean Age	St. Dev. Age	% Female	% Married	n
1	1973	37.0	(10.28)	50.0	80.5	3008
2	1981	44.3	(10.05)	50.1	82.9	2921
3	1985	48.3	(9.99)	51.1	83.0	2849
4	1988	51.5	(9.99)	51.5	80.6	2796
5	1992	55.0	(9.83)	52.1	79.9	2709
6	1996	58.6	(9.69)	52.7	77.2	2613
7	1999	61.5	(9.58)	53.1	74.7	2565

Ages in the sample range from 13 in exam 1 to 88 in exam 7.

has been eliminated, the number of individuals at each exam reflects only those that have survived. Over the health exams, the sample becomes slightly more weighted toward female and non-married individuals. Table 4.2 also shows the great variability in ages across the sample. At the first exam, there are individuals who are as old as the average age at the final exam. Indeed, over the entire sample, ages range from 13 to 88. Finally, because of the small ethnic minority population in Framingham, Massachusetts, particularly at the onset of the Original Cohort, for reasons of confidentiality, there is no racial variation in the sample and

all individuals are white.

Table 4.3 gives sample percentages of the maximum number of years of education by category. The sample reflects a rather well educated cohort for the time period as nearly 89% of the sample has a high school degree or better.

Table 4.3: Education

Education Years	**% of Sample**
0-4	3.2%
5-8	1.0
9-11	6.1
12	32.8
13-16	43.2
17+	13.8

$N = 3008$. Percentages reflect highest attained level of education.

Table 4.4 breaks down the sample by smoking prevalence over exams. Over the seven exams, the sample smoking prevalence drops from roughly 41% to 11%. Interestingly, at the first exam, smoking prevelence in the sample is roughly consistent with that of the United States average prevelence (37% of Americans smoked in 1973).[3] However, by the final exam, the sample percentage of smokers has decreased to roughly 11% whereas the national average fell to 23.3%. The sample is also clearly older than the general population by the seventh exam.[4]

The health markers that I consider are identified by the Framingham Heart Study as those that are most important in determining an individual's general 10-year risk for cardiovascular disease (Wilson *et al.*, 1998; D'Agostino *et al.*, 2008).[5] These include systolic and diastolic blood pressure, total and high-density lipoprotein cholesterol, and diabetes. I do not consider BMI because, as noted in D'Agostino *et al.* (2008), total cholesterol is a better risk indicator

[3] See Chapter 5 for a complete comparison between the Framingham Heart Study participants and the overall United States prevalence of smoking.

[4] Centers for Disease Control and Prevention: http://www.cdc.gov/tobacco/basic_information/index.htm

[5] The Framingham Heart Survey: Risk Score Profiles:http://www.framinghamheartstudy.org/risk/index.html

Table 4.4: Smoking Behavior by Exam

Exam	Nonsmokers	Light Smokers $\leq$ 1 Pack/Day	Heavy Smokers $>$ 1 Pack/Day
1	59.0%	26.7%	14.3%
2	61.3	24.4	14.2
3	77.2	14.3	8.5
4	81.2	12.8	6.0
5	85.2	11.0	3.9
6	87.9	9.5	2.6
7	88.9	8.7	2.5

when available.[6] Table 4.5 provides summary statistics of these variables at each health exam.

Recall that the theoretical model specifies a dichotomous variable for the presence of a

Table 4.5: Health Markers by Exam

Exam	Systolic Blood Pressure	Diastolic Blood Pressure	High-Density Lipoprotein (HDL)	Total Cholesterol	Diabetes
1	122.3	79.0	50.8	197.1	1.8%
	(16.1)	(10.7)	(15.0)	(39.4)	
2	122.3	78.2	48.6	203.9	2.6
	(16.5)	(9.8)	(13.6)	(39.1)	
3	123.7	79.0	51.1	212.2	3.5
	(16.7)	(9.6)	(14.8)	(41.2)	
4	126.8	79.1	49.9	207.5	4.8
	(18.8)	(10.0)	(14.8)	(38.5)	
5	126.4	74.5	49.9	205.6	6.9
	(18.8)	(10.1)	(15.2)	(36.5)	
6	128.2	75.3	51.0	206.0	9.6
	(18.4)	(9.5)	(16.0)	(37.7)	
7	127.2	73.8	53.3	200.6	11.1
	(18.7)	(9.7)	(16.8)	(36.6)	

Standard deviations are in parenthesis.

chronic condition. The decision to aggregate the data to a dichotomous, chronic condition or not, level stems from the computation burdens of estimating additional parameters in the structural model. As in the theoretical model, I assume that upon transiting to a chronic health

[6]Of course, evidence has shown that body mass index (BMI) and smoking are inversely related. However, assessing the potential weight gain associated with smoking cessation is beyond the scope of this dissertation. Furthermore, BMI is a very noisy measure of health, and the Framingham Heart Study 10-year cardiovascular risk calculator states that, when available, total cholesterol should replace BMI when possible.

state, an individual remains in that state for life. Table 4.6 shows the percentage of the sample living with a chronic condition at each health exam. The incidence of new chronic conditions is in column 3 of table 4.6. I consider an individual to have a chronic condition if any of a

Table 4.6: Chronic Health by Exam

Exam	**Chronic Condition at Exam**	**Newly Chronically Ill at Exam**
1	0.2%	0.0%
2	4.0	3.8
3	7.0	3.4
4	9.6	3.1
5	12.2	3.9
6	16.0	5.0
7	20.3	4.8

wide variety of cardiovascular diseases and cancers are present. Unfortunately, this definition does exclude some smoking related chronic conditions such as chronic obstructive pulmonary disease. This definition is a limitation of the data. The panel nature of the Framingham Heart Study focuses on heart related conditions and cancers.[7] Table 4.7 shows the prevalence of the five types of heart conditions considered. Myocardial Infarction, or heart attack, makes up the largest percentage of heart conditions. Various cancers constitute the remaining 28.3% of the defined chronic conditions. While a large number of cancers are thought to be related to cigarette smoking, I include all types of cancer as a chronic condition. Finally, because data

Table 4.7: Chronic Health Conditions

Condition	**Number**	**Percentage**
Heart Conditions	1980	71.7%
Myocardial Infarction	559	28.2%
Other Coronary Heart Disease	488	24.6
Cerebrovascular Accident	323	16.3
Intermittent Claudation	182	9.2
Congestive Heart Failure	428	21.6
Cancers	781	28.3%

[7] See Chapter 10 for a discussion of the limitations of the chronic health definition.

are available for any chronic conditions that occurred prior to entering the sample, as well as, the date of a chronic condition after entering the sample, Figure 4.1 shows the cumulative empirical distribution of individuals in the chronic health state by age. Note that, because the sample is right censored, and because of death, the data become noisy for ages beyond 75.

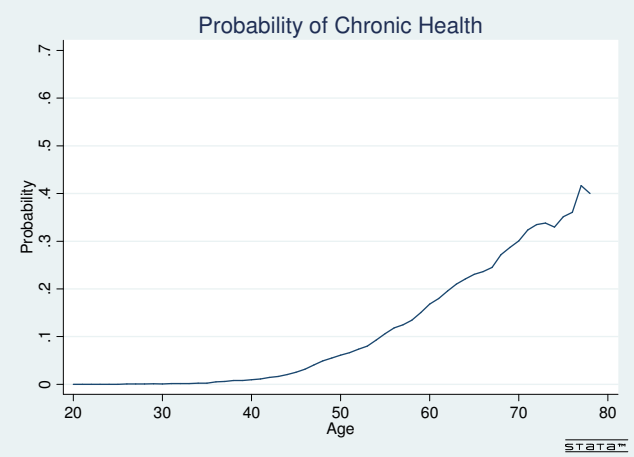


Figure 4.1: Sample Chronic Health Condition Probabilities by Age

Chapter 5

Empirical Implementation

In estimating the model described in Chapter 3 with the data described in Chapter 4, there are four hurdles. The first main hurdle lies in the timing of the health exams. While the data contain only seven exams over a 40-year period, the theoretical model is based on a yearly decision making process over a finite time horizon. As explained below, I exploit retrospective questions in the data to construct a dataset that mirrors the timing of the model. This process generates yearly data for all variables in the theoretical model *except* health marker data for years in which no health exam took place. To overcome this problem, I use predictions from solution to the model to integrate over "off years" as well as to explain the initial condition for each individual. Given the expanded dataset, the second hurdle is that state variables A_{it} and R_{it} must be constructed from the Framingham Heart Study data in such a way as to capture an individual's smoking history and health markers, respectively. For each of these variables, I employ principal component analysis in a method similar to that of Sickles and Williams (2008).[1] The third hurdle is the identification of the model parameters. As I describe below, variation in the timing of health exams across individuals helps to identify the model. Finally, the last hurdle involves modeling permanent unobserved heterogeneity and resolving the initial conditions problems. This chapter ends with a detailed discussing of my solution method, likelihood function construction, and maximization routine.

[1] See Chapter 2 for a discussion of Sickles and Williams (2008).

5.1 Health Exam Timing

While I observe individuals at only seven health exams over a 40-year period, the theoretical model is based on a yearly decision making process. To reconcile this difference, I proceed in the following steps. First, in solution to the model, I specify the final period, T, to be at age 100. That is, the probability of death at the end of period T equals one. The yearly model is then solved recursively back to age 7, at which point I assume that all individuals have a smoking stock of zero (i.e. $A_{i7} = 0$, $\forall i$). This process is described in detail below. Second, the data from chapter 4 are expanded based upon retrospective questions. With the exception of the health marker information needed to construct the health marker index, R_{it}, data are available to construct a yearly dataset from age seven until an individual either dies or completes his or her seventh exam. Data in years prior to an individual's first exam were constructed based on questions at the first and second exams that asked, if applicable, the first age at which one started smoking and the age at which one stopped smoking. For later years in between health exam years, smoking data were imputed based on history and adjacent health exam data. Specific dates are available in the data for any chronic health and mortality events.

Figure 5.1 shows the sample probabilities for each smoking choice by age. Because figure

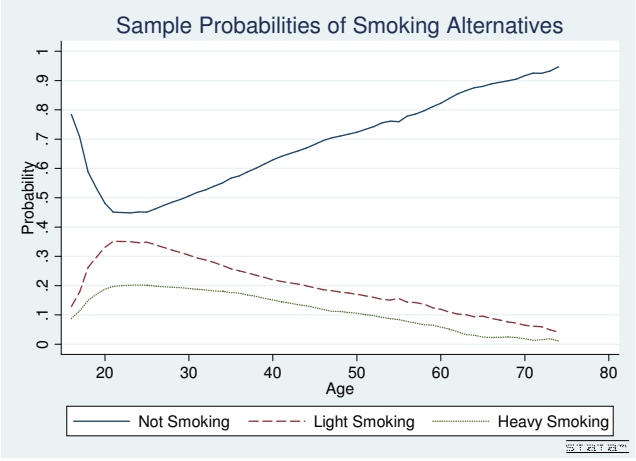


Figure 5.1: Sample Choice Probabilities by Age

5.1 reports smoking percentages by age, for older ages, the data become noisy because of either

right-censoring or death. Of the 3008 observations considered, 15%, or 464 individuals, leave the sample through death.

Table 5.1 presents age and gender specific smoking rates by year from the Framingham Heart Study and the United States national averages for white Americans.[2] As my sample of individuals ages, certain age groups drop out of the sample; however, almost the entire sample is present in 1974 and 1985. Interestingly, while generally both men and women in my sample are consistently below the United States smoking averages, Framingham Heart Study women tend to smoke more in 1965 and 1974 than average. Furthermore, following some age cohorts

Table 5.1: Framingham Heart Study and United States Smoking Prevalence by Age and Year

	1965		**1974**		**1985**		**1995**	
	FHS	U.S.	FHS	U.S.	FHS	U.S.	FHS	U.S.
White Males, Age								
18-24	53.2%	53.0	39.3	40.8	-	28.4	-	28.4
25-34	58.9	60.1	43.4	49.5	29.0	37.3	-	29.9
35-44	54.2	57.3	37.9	50.1	29.9	36.6	15.0	31.2
45-64	53.2	51.3	37.2	41.2	22.7	32.1	13.9	26.3
65+	-	27.7	-	24.3	16.9	18.9	6.1	14.1
N	1487		1496		1345		1086	
White Females, Age								
18-24	50.0	38.4	40.0	34.0	-	31.8	-	24.9
25-34	50.3	43.4	41.2	38.6	22.5	32.0	-	27.3
35-44	42.5	43.9	35.6	39.3	27.0	31.0	17.2	27.0
45-64	45.3	32.7	35.8	33.0	22.5	29.7	14.1	24.3
65+	-	9.8	-	12.3	23.0	13.3	9.7	11.7
N	1496		1502		1364		1206	

over time shows that, while Framingham Heart Study men aged 18-24 in 1965 appear to quit smoking faster than the national average, Framingham Heart Study women aged 18-24 in 1965 smoke more in 1965 and 1974 than the nation average, but then appear to quit more rapidly in 1985 and 1995.[3] Table 5.1 puts the Framingham Heart Study participants in the context of the United States averages. Initially, individuals in my sample, corrected for age, race, and

[2]Data are from http://www.cdc.gov/nchs/hus/contents2010.htm#table058.

[3]To make these comparisons, examine the prevalence of smoking for 18-24 years olds in 1965 by comparing the difference between Framingham Heart Study individuals and United States averages. Next, because these individuals are all between 25-34 years old in 1974, compare these averages.

gender, look similar in terms of smoking prevalence to the rest of the United States. However, by 1995, a smaller percentage of my sample smokes than average.

Table 5.2 shows general smoking summary statistics from the expanded data. On average, if an individual ever smokes, he/she starts just before age 20, although the median age is 18. Table 5.3 reports smoking behavior transitions around health exams and chronic health

Table 5.2: Smoking Summary Statistics

	Mean (Median)	**S.D.**	**Min**	**Max**
First Age Smoking	19.57 (18)	7.45	7	67
(Conditional on Ever Smoking)				
Total Years Smoking	24.78 (24)	14.07	1	68
(Conditional on Ever Smoking)				
Tenure Smoking (Years)	21.13 (19)	14.85	1	68
(Conditional on Ever Smoking)				
Last Age Smoking	44.76 (45)	12.84	13	76
(Conditional on Ever Smoking and Quitting)				

shocks, as well as the overall average transitions. Conditional upon an individual's smoking behavior one period prior to each event, the table reports percentages in each smoking option one and three years after the event. For example, of those individuals smoking heavily one

Table 5.3: Observed Transitions

	Behavior One Period Post			Behavior Three Periods Post		
Behavior One Period Prior	Not Smoking	Light Smoking	Heavy Smoking	Not Smoking	Light Smoking	Heavy Smoking
Overall Transitions						
Not Smoking	95.86%	2.93%	1.22%	91.72%	5.76%	2.51%
Light Smoking	9.21	89.29	1.51	17.35	79.81	2.84
Heavy Smoking	5.66	2.78	91.56	11.12	5.31	83.58
Transitions Around Health Exams						
Not Smoking	96.38	3.20	0.42	97.00	2.55	0.45
Light Smoking	21.40	70.29	8.30	29.45	63.50	7.06
Heavy Smoking	14.50	16.47	69.03	20.62	16.48	62.90
Transitions Around Chronic Health Shocks						
Not Smoking	98.94	0.89	0.18	97.18	2.59	0.24
Light Smoking	24.78	73.45	1.77	39.36	54.26	6.38
Heavy Smoking	21.43	8.33	70.24	39.71	13.24	47.06

period prior to a chronic health shock, 70.24% continued to smoke heavily one period after the shock and 47.06% were still smoking heavily three periods after the shock. In both event cases, more individuals had quit smoking three years post as compared to one year post, but

considerably more had quit three years after an event than the baseline set of transitions. The table provides at least antidotal evidence that each event (health exams and chronic health shocks) alters smoking behavior in the sense that the magnitudes of the transitions, both one and three years post, are larger than the overall transitions. I compare these descriptive statistics with predictions from the model in chapter 7.

5.2 Continuous State Variable Construction

For the model to both remain computationally tractable and be consistent with the assumption of conjugate distributions, I need continuous, scalar representations of both an individual's smoking history and her health markers. My solution is to employ principal component analysis (PCA) in the construction of each variable. PCA is a nonparametric technique that summarizes the total variation in a set of variables into an ordered set of continuous, scalar principal components. The first principal component is constructed as a linear combination of data and of factor loadings from the highest eigenvalue eigenvector from an eigenvector decomposition of the variables' correlation matrix.[4] The trade-off with PCA is both completeness and interpretation. Only considering the first principal component implies that any remaining variation in the data (i.e. the second, third, fourth, etc. principal components) is lost. Furthermore, because the weights used to construct the index are derived only from the correlation between the variables themselves, the relative magnitudes of the weights may come into question when predicting an outcome of interest (e.g., mortality). In the context of most structural models, however, reducing the dimension of the data is clearly adventageous. I will discuss these issues for both the smoking stock and health marker index construction, as well as alternative specifications, below.

In constructing R_{it}, the health marker index, I use PCA with the following (standardized) variables: systolic blood pressure, diastolic blood pressure, total cholesterol, high-density

[4]For example, given a set of k variables, employing PCA will yield k principal components. If, however, the first two principal components account for 70% of the total variation in the k variables, and $k > 2$, the researcher may find it adventagous to only use the first two principal components as data.

lipoprotein (HDL) cholesterol, and a diabetes dummy.[5] As noted in Chapter 4 and Table 4.5, these health markers are identified by the Framingham Heart Study as significant predictors for an individual's general 10-year risk of cardiovascular disease.[6] The first principal component of these variables explains approximately 33% of the total variation. Unfortunately, this implies that two-thirds of the variation in health markers is being lost. However, I now have a continuous index of health markers. I see two main justifications for using the first principal component as my measure for the health markers. First, the theory places no restriction on the amount of information that R_{it} must convey, only that it conveys some information. Any computationally tractable definition of R_{it} will have to be an approximation. That I can explain a third of the variation in the variables that the medical literature view as signficant will at least inform to some degree. Second, most papers that use PCA use first principal components that explain between $20\% - 40\%$ of the total variation.[7]

To provide intuition as to the weights used to create the health index, Table 5.4 presents results from an OLS regression of the health index on the above health markers. The regression is run without a constant. Point estimates correspond to the eigenvector values from the first principal component for the corresponding variables. In this context, the continuous health index can be interpreted as a measure of bad health (i.e. higher values of the index imply worse overall health). Note in Table 5.4 that only HDL, or "good" cholesterol, negatively affects the health index.

As discussed above, the smoking stock summarizes all past smoking decisions prior to period t. Again using PCA, I define the index A_{it} as the first principal component of the following four standardized variables: total number of years smoking at time t (experience),

[5]PCA is most effective when there exists significant correlation between the variables. As one might expect, the correlation between these health markers is high.

[6]While influential in predicting cardiovascular disease, there is no evidence that suggests that these health markers predict different forms of cancer. In the context of the model, a summary of these health markers will have less predictive power on the chronic health state if that state is defined as an aggregation of cardiovascular disease and cancer. However, these markers still provide an overall assessment of an individual's health. The extent to which these markers may influence smoking behavior through the Bayesian updating process is an empirical question.

[7]In the context of socioeconomic indices, see Vyas and Kumaranayake (2006) for a good overview of PCA.

Table 5.4: Health Index Regression

Variable	Coefficient		(Std. Err.)
Systolic Blood Pressure	0.657	***	(0.000)
Diastolic Blood Pressure	0.637	***	(0.000)
Total Cholesterol	0.306	***	(0.000)
High-Density Lipoprotein (HDL)	-0.177	***	(0.000)
Diabetes	0.193	***	(0.000)

Significance Levels: ***1% Level, **5% Level, *10% Level

number of years smoking at time t since last year not smoking (tenure), number of years at time t not smoking since last year smoking (cessation), and the intensity of smoking in the previous period, $t-1$.[8] I term these variables experience, duration, cessation, and intensity respectively. Table 5.5 gives sample averages by exam of the number of years of duration, tenure and cessation. The first principal component explains nearly 52% of the total variation

Table 5.5: Smoking History by Exam

Exam	Experience	Duration	Cessation
1	10.6	7.3	1.9
2	13.0	7.9	4.1
3	14.1	7.6	4.8
4	14.7	6.2	6.1
5	15.1	5.3	7.9
6	15.1	4.4	9.8
7	15.3	3.8	11.3

Values are in years.

in these four variables.

To aid in interpretation of both the resulting smoking stock and the associated parameters to be estimated, I normalize the smoking stock as follows. First, I run PCA on just those with some smoking history. That is, individuals with any observed or reported past smoking in each period are included in the PCA. For example, if an individual takes her first exam at age 18 and begins smoking at age 22, all observations from this individual after age 22 are included in the PCA, whereas observations prior to 22 are not included. Second, I shift the distribution of the resulting index such that the person with the lowest value has a stock approximately

[8]Intensity is measured as the average number of cigarettes per day. Each of these smoking variables is measured as the value entering the examination.

equal to zero. Finally, for individuals with no smoking history, I assign a stock value of zero. Table 5.6 reports the results of a regression of the smoking stock index on the four variables of interest (excluding a constant). Point estimates correspond to the eigenvector values from the first principal component for the corresponding variables.

Table 5.6: Smoking Stock Regression

Variable	**Coefficient**	**(Std. Err.)**
Experience	0.309***	(0.000)
Duration	0.589***	(0.000)
Cessation	-0.517***	(0.000)
Intensity	0.540***	(0.000)

Significance Levels: ***1% Level, **5% Level, *10% Level

Notice that while experience, duration, and smoking intensity of an individual all increase the stock index, cessation from smoking decreases the stock through the depreciation parameter, δ_1. I therefore interpret higher values of the index as more accumulated smoking stock capital.

5.3 Identification

As is discussed below, the parameters of the model are estimated via simulated maximum likelihood. Generally, the structural model is identified from the variation in the timing of the health exams. While the number of years between health exams does not directly affect health[9], observationally equivalent individuals with different time gaps between exams may select different smoking patterns. The different smoking patterns may arise because different time gaps will induce variation in the belief distribution across individuals. For example, if, after completing their second health exam at the same time, one individual receives her next exam in three years while another individual receives her next exam in five years, the data may show different smoking patterns during the two years in which the first individual had a different set of beliefs. Table 5.7 shows variation in the number of years between exams (the vertical tab) for each exam (the horizontal tab). While the gap between the first and second

[9]The considerable time lag between exams is because the Framingham Heart Study administers health exams in time "windows". However, there is no evidence to suggest that those with worse health markers select into smaller time gaps between exams.

Table 5.7: Variation in Exam Timing

Years	**Exam**					
Between Exams	**2**	**3**	**4**	**5**	**6**	**7**
1	0	0	4	0	0	103
2	0	27	63	36	49	718
3	0	141	1659	358	326	958
4	0	1951	1000	2242	2019	682
5	6	657	57	55	184	90
6	27	65	12	16	28	13
7	1500	8	1	2	7	1
8	1302	0	0	0	0	0
9	64	0	0	0	0	0
10	22	0	0	0	0	0
Total	2921	2849	2796	2709	2613	2565

exam is clearly the longest, and the time gap shrinks at later exams, each exam exhibits considerable variation across individuals in the number of years to be administered. The following set of parameters are to be estimated.

Utility Parameters:	$\Theta_U = \{\alpha_{0h}, \ldots, \alpha_{11h}\}_{h=0}^{1}$
Health Transition Parameters:	$\Theta_H = \{\lambda_0, \ldots, \lambda_{10}\}$
Death Transition Parameters:	$\Theta_M = \{\omega_0, \ldots, \omega_{12}\}$
Smoking Stock Parameters:	$\Theta_A = \{\delta_1, \delta_2, \delta_3, \sigma_\eta\}$
Learning and Risk Parameters:	$\Theta_R = \left\{\overline{\theta}, \sigma_\theta, \sigma_\nu, \phi, \zeta\right\}$
Factor Loadings:	$\Theta_\rho = \left\{\left\{\left\{\rho^{Uhd}\right\}_{h=0}^{1}\right\}_{d=0}^{2}, \rho^H, \rho^M, \rho^R, \rho^A\right\}$

Additionally, I estimate the probability weights of the mass points for the discretized distribution of the permanent unobserved heterogeneity, μ. Let $\Theta = \{\Theta_U, \Theta_H, \Theta_M, \Theta_A, \Theta_R, \Theta_\rho\}$. In order to identify the preference parameters, I normalize the utility of death to be zero. Relative to this normalization, identification of the preference parameters comes mainly from variation in smoking behavior and health and death transitions over time. For example, different smoking choices across the smoking stock, health marker index, and age levels identifies the interaction preference parameters. Furthermore, the withdrawal parameter, $\alpha_{9\cdot}$, is identified off of variation in the choices of individuals after a period in which an individual quits.

Thus, conditional upon having smoked in the previous period, both the reinforcement, $\alpha_{2\cdot}$ and $\alpha_{6\cdot}$, and withdrawal effects, $\alpha_{9\cdot}$, encourage current period smoking. However, withdrawal is separately identified from the reinforcement effects because, while the smoking stock variable depreciates at rate δ_1 following cessation, the utility cost paid from withdrawal only lasts one period. Finally, the direct impacts of the stock on utility, $\alpha_{10\cdot}$ and $\alpha_{11\cdot}$, reflect tolerance in smoking and are identified by individuals that progress from light to heavy smoking.

In the absence of subjective expectation data, the structure of the model is needed to identify the presence of learning. Mira (2007) notes that learning can no more be identified than can rational behavior in the sense that, the model assumes that behavior (learning) follows from the defined structure. If, however, the prior distribution of beliefs is proved to be degenerate (i.e., if the null hypothesis that $\sigma_\theta = 0$ is not rejected), then the results would suggest an absence of learning. The identification strategy of the specific learning parameters is therefore quite subtle. While identification of $\overline{\theta}$ comes from variation in the smoking stock and health marker index, variation in smoking by individuals over time identifies σ_θ (Crawford and Shum, 2005). If, indeed, individuals are learning over time, choices at the end of the time frame relative to the beginning should better reflect an individual's true match value, θ_i. An additional source of variation that helps to identify the learning parameters is the variation across individuals in the timing of health exams. There exists considerable variation in the number of years between exams across individuals; thus, two similar individuals that receive health information at different frequencies may develop different smoking patterns. Because of the assumption of conjugate normal distributions, identifying the mean and variance of θ_i, in addition to the variance of ν, which is identified from the health marker index transition equation, is sufficient to characterize the learning process. Finally, the identification of chronic health and death transition parameters comes from variation in the state variables and the observed incidence of chronic health and death.

5.4 Permanent Unobserved Heterogeneity and Initial Conditions

Permanent unobserved heterogeneity enters the model in a linear fashion through the μ term and the associated factor loadings. The factor loadings allow for a different effect of the unobserved μ term everywhere it enters. Rather than placing a distributional assumption on the underlying unobserved heterogeneity, I approximate its distribution with a step function and estimate the factor loadings and mass point probabilities with other parameters in the model (Heckman and Singer, 1984). This discrete factor method has been shown to approximate both Gaussian and non-Gaussian distributions well (Mroz, 1999).

I first observe individuals at various points in their life cycle (i.e., different ages at the first health exam) and with a variety of health histories. Failing to properly model these histories would lead to an initial conditions problem. Furthermore, the initial conditions problem may lead to an issue of dynamic selection into smoking behaviors. That is, individuals in some permanently lower (unobserved) health state may select into smoking. However, solution to the model generates individual probabilities of choice behavior and health/death transitions for all ages beginning at age seven. Recall that data exist for all smoking, chronic health, and death events from age seven until either death or the final health exam (exam 7 in the data). At age seven, I assume that each individual has a smoking stock of zero and has no chronic health problems. The only remaining initial condition is the initial health marker index upon entering the sample. Using the model, I can simulate a health marker index for each period from age seven until the first observed health exam. Hence, I use the model to generate probabilities of an individual's health history when they are first observed in the sample (Khwaja, 2010). Individuals enter the sample aged between 13 and 62 years. At age seven, I assume that the lagged value of the health marker index is in the 90th percentile (e.g., good health) of each health marker that is used to construct the index. I then use the weights from the principle component analysis to construct the lagged value. Recall that the simulated health marker index is scaled by demographic characteristics, X_{it}, as well as the unobserved heterogeneity, μ, term and its factor loading. Furthermore, individual variation in the data at the first exam (the initial condition) helps to identify parameters of the model.

5.5 Solution

The computational hurdle in calculating the conditional choice probabilities in equation 3.9 is to solve for the integrated Bellman (EMAX) equation in equation 3.8. Technically, the EMAX equation must be solved for all possible points s in the state space S_{it}. However, given the long time frame of the model and the mixed discrete/continuous nature of the state space, I employ a variant of the Keane and Wolpin (1994) value function interpolation method for approximating the value function. This method amounts to drawing from the state space, calculating the resulting EMAX function for each draw, and interpolating the EMAX function for all other points. The end goal of this procedure is to generate choice probabilities for each individual i, in each time period t, conditional on the unobserved heterogeneity μ and a trial set of the parameters, to enter the likelihood function. My iterative solution method proceeds in two main steps: model simulation and individual specific solution. While the first solution step yields value function regression coefficients from the simulated model, the second step uses these coefficients to calculate the conditional choice probabilities, health marker and smoking stock densities, and chronic health and mortality transition probabilities.

The first step of the solution method is to solve the model for a group of simulated individuals. The goal of this step is to generate a set of regression coefficients that map from the state space to the value function. Because the time horizon is finite ($T = 100$), I can solve the model using backwards induction and I avoid iterating on the value function itself. Starting in the final period T, I draw n state vectors and sequences of past smoking behavior $\mathbf{D}_{iT-1} = \{d_{i1}, \ldots, d_{iT-1}\}$.[10] Each of these n draws represents one simulated individual. For each of the n draws, I construct the main equations of the model for period T. Note that because the probability of death at the end of period T equals one, each of the choice specific value functions in period T simply equals the current period utility from the smoking alternative. Next, I posit a relationship between the n calculated value functions and a set of regressors. The regressors include the drawn state variables in addition to interaction and higher-order terms. I then run the regression and generate coefficients that are specific to time

[10] In practice I set $n = 100$.

period T. Next, I repeat the above steps for period $T-1$. When calculating the expected value function in period $T-1$, I use the regression coefficients from period T to approximate the expected future value function. I repeat the above process for all periods back to age seven, $t = 7$; that is, I solve the model for all ages between 7 and 100.

The first stage process is conditional on three factors. First, I conduct the simulation above for each possible age at which an individual may have taken her first health exam.[11] Second, I discretized the support of the unobserved heterogeneity distribution into K points.[12] For each point k of μ, in addition to each age at the initial health exam, I conduct the above simulation. Thus, I have a full set of value function regression coefficients (from age 7 to 100) for each age at initial health exam and for each unobserved type, μ. Finally, the value function regression coefficients are also conditional on the trial set of parameters used to solve the model.

The second main step, conditional on the same trial set of parameters and using the above regression coefficients, involves solving the model for each individual. For each individual, I solve the model backwards from age 100 to generate conditional choice probabilities, health marker and smoking stock densities, and chronic health and mortality transition probabilities. This process is complicated by the fact that I only have data for the health marker index in some periods.

In each period in which individual i undergoes a health exam, he/she must forecast the future evolution of the state variables and the resulting values associated with all current and future smoking decisions. Luckily however, because the value function regression coefficients approximate the next period value function, I must only construct the expected value of the next period state variables conditional on the current period smoking decision. Because the chronic health and mortality logit probabilities have closed-form expressions, assuming rational expectations makes the next period chronic health and mortality transition expectations

[11] In the data, the ages range from 13 to 62. As noted above, there exists great variation in the data in the timing of the health exams. However, in the simulation, regardless of age at the initial health exam, I use the average number of years between exams to avoid having to simulate the model for all possible combinations of exam sequences.

[12] In practice I set $K = 3$.

straightforward. To forecast the smoking stock and health marker index values one period forward, I use a Monte Carlo method. Conditional on each draw of the Monte Carlo simulator, I construct all other probabilities in the model. The average of these probabilities then enters the likelihood function.

For time periods in which no health exam was taken, in addition to integrating over the future values of the smoking stock and health marker index, I must also integrate over the *current* period value of the health marker index. In this case, I use the exact same method as integrating over the future health marker index, only using different draws (from those used to integrate over the future term) of the i.i.d. error term ν. All other probabilities in the model are constructed conditional on the drawn value of the current period health marker index and averaged. For a more formal explanation of this method, I now present the construction of the simulated likelihood function.

5.6 Likelihood Function

Consider first the contribution of individual i to the likelihood function. Given that $\eta_{it} \sim \mathcal{N}(0, \sigma_\eta^2)$ and $\nu_{it} \sim \mathcal{N}(0, \sigma_\nu^2)$, I can express the probability density functions of A_{it} and R_{it} respectively as:[13]

$$\Lambda_t = f(\eta_{it}|A_{it-1}, d_{it-1}, \mu, \rho_A, \Theta_A) = \frac{1}{\sigma_\eta}\phi\big([\log A_{it} - \delta_1 \log A_{it-1} - \delta_2 d_{t-1} - \rho_A \mu]/\sigma_\eta\big) \tag{5.1}$$

and

$$\Omega_t = g(\nu_{it}|X_{it}, \kappa_{it}, \mu, \rho_R, \Theta_R) = \frac{1}{\sigma_\nu}\phi\big([R_{it} - \zeta R_{it-1} - X_{it}\phi - \kappa_{it} - \rho^R \mu]/\sigma_\nu\big) \tag{5.2}$$

where $\phi(\cdot)$ is the standard normal distribution. Recall, however, that the health marker index, R_{it}, and only the health marker index, is unobserved in periods in which a health exam was

[13]Note that the i subscript has been dropped from the permanent unobserved component, μ. As is shown below, in the empirical model, the distribution of μ has been discretized to K points of support and is integrated out of the likelihood function.

not taken.[14] I must, therefore, integrate over R_{it} in all periods with no health exam. For ease of exposition, define the dummy y as follows:

$$y_{it} = \left\{ \begin{array}{lll} 1 & \text{if} & \text{An exam was taken in year t} \\ 0 & \text{if} & \text{No exam was taken in t} \end{array} \right\}$$

Define $Z_{it}^{y=1}|\mu$ as individual $i's$ likelihood contribution in period t when $y_{it} = 1$ and conditional on unobserved heterogeneity term μ:

$$Z_{it}^{y=1}|\mu = \prod_{d=0}^{2} \left(p(d_{it} = d|s_{it}, \mu) * \Lambda_t * \Omega_t * \prod_{h=0}^{1} \left(\pi_{t+1}^{dh}|\mu \right)^{\mathbf{1}[H_{it+1}=h]} * \prod_{m=0}^{1} \left(\varsigma_{t+1}^{dm}|\mu \right)^{\mathbf{1}[M_{it+1}=m]} \right)^{\mathbf{1}[d_{it}=d]}. \quad (5.3)$$

Here, π_{t+1}^{h} represents the probability of transiting to health state h in period $t+1$ and ς_{t+1}^{m} is the probability of transiting to death state m in period $t+1$. Unless a health exam was taken in the period directly before t, the lagged value of the health marker index in equation 13 is unobserved. In practice, I use the expected health marker index given the model parameters as the lagged value. In periods in which $y_{it} = 0$, define the expected health marker index, conditional on the model parameters as:

$$\widetilde{R}_{it} = \mathrm{E}_{\nu}(R_{it}|\Theta_R, S_{it}, \mu) \quad (5.4)$$

Here, the expectation operator is taken over the i.i.d. noise term, ν. Other probabilities in the model are conditional on $\widetilde{R}_{it}$ for years in which $y_{it} = 0$.[15] In the period directly after a health exam, the lagged value of the health marker index (i.e., from the exam and not the simulated term) is used in the construction of $\widetilde{R}_{it}$. Therefore, define $Z_{it}^{y=0}|\mu$ as individual $i's$ likelihood

[14] All right-hand side terms in the health marker equation *are* observed in these "off" years due to retrospective questions and/or imputation with the exception of the lagged value of the health marker index when the previous period did not contain a health exam.

[15] In practice, I numerically integrate over ν_{it}. For each draw of ν_{it}, all other probabilities in the model are constructed. The resulting probabilities are then averaged over the draws.

contribution in period t when $y_{it} = 0$:

$$\prod_{d=0}^{2}\left(p(d_{it}=d|s_{it},\widetilde{R}_{it},\mu)*\Lambda_{t|\widetilde{R}_{it}}*\prod_{h=0}^{1}\left(\pi_{t+1}^{dh}|\widetilde{R}_{it},\mu\right)^{\mathbf{1}[H_{it+1}=h]}*\prod_{m=0}^{1}\left(\varsigma_{t+1}^{dm}|\widetilde{R}_{it},\mu\right)^{\mathbf{1}[M_{it+1}=m]}\right)^{\mathbf{1}[d_{it}=d]}. \tag{5.5}$$

The total conditional (on μ) likelihood contribution from individual i for all time periods $7,\ldots,T_i$, where T_i is either the period of an individual's death or their final exam, is:

$$L_i(\Theta|\mu)=\prod_{t=7}^{T_i}\Big[\prod_{y=0}^{1}\Big(Z_{it}^{y}|\mu\Big)^{\mathbf{1}[Y_{it}=y]}\Big]. \tag{5.6}$$

Because of the discretized distribution of the unobserved heterogeneity, each individual's *unconditional* contribution will be a finite mixture of likelihoods. Given K points of support in the estimated distribution of μ, the unconditional likelihood function contribution for individual i is:

$$L_i(\Theta)=\sum_{k=1}^{K}\xi_k L_i(\Theta|\mu_k). \tag{5.7}$$

Where ξ_k is the estimated probability weight placed on mass point k. The full sample log-likelihood function is:

$$L(\Theta)=\Big[\sum_{i=1}^{N} logL_i(\Theta)\Big]. \tag{5.8}$$

The parameter estimates in Θ are estimated via a nested solution method (Rust, 1987). The inner algorithm solves the dynamic model for each individual conditional on a given set of parameters and for all mass points of the unobserved heterogeneity distribution. Using the resulting probabilities, the outer algorithm calculates the unconditional likelihood function, $L(\Theta)$, and attempts to improve the likelihood value via a BHHH gradient method. The BHHH method is standard in estimating dynamic structural models because, as opposed to traditional gradient methods such as Newton-Raphson that explicitly construct the Hessian matrix of the likelihood function. BHHH approximates the Hessian by exploiting the fact that the likelihood function ($L(\Theta)$) is the sum of individual log-likelihood contributions. Calculating the second derivatives of the likelihood function would be computationally infeasible for nearly all dynamic structural models. At the parameters that maximize the log-likelihood function,

however, the average outer-product over individuals is the covariance matrix of the scores of the sample. Furthermore, at the true parameters, the covariance matrix of the scores is equal to minus the expected Hessian matrix (Train, 2009). The total process continues, updating parameters at each iteration, until the likelihood function is maximized. I assume convergence at the maximum of the likelihood function when the percentage change of the likelihood value over an iteration is at or below 0.000001. The model is solved and estimated using MPI parallel processing techniques for Fortran 90 code.

Chapter 6

Results

6.1 Parameter Estimates

Table 6.1 reports the main parameter results. The estimated utility constants, α_{00} and α_{01}, for the absence of a chronic health condition and a chronic health condition respectively, are quite intuitive given that the utility of death has been normalized to zero. The total marginal utility of current period light and heavy smoking is a function of $\alpha_1 \ldots \alpha_8$.. A key component of rational addiction theory, indeed the defining feature of an addictive good under rational addiction, is that past consumption reinforces current consumption. That is, the marginal utility of smoking is increasing in the smoking stock. My results are consistent with this adjacent complementarity defined in Becker and Murphy (1988). In the absence of a chronic illness, both light and heavy smoking are found to be reinforcing (i.e., α_{20}, $\alpha_{60}>0$). Indeed, I find that heavy smoking is much more "reinforcing" than light smoking. My results also suggest that the marginal utility of light smoking in the absence of a chronic condition is invariant to the health marker index but increasing in age. Interestingly, the marginal utility of heavy smoking is decreasing in the health marker index and invariant to age when free of a chronic condition; however, when chronically ill, the marginal utility of heavy smoking is increasing in the health markers ($\alpha_{71} = 0.001$) and decreasing in age. Withdrawal from smoking, (i.e., smoking in period $t-1$ and not smoking in period t) is negative for all health states and larger in magnitude when free of a chronic illness. The withdrawal effect, in addition to the strong reinforcement effect, both drive smokers to continue smoking. Finally, the tolerance

effect ($\alpha_{10\cdot}$) flips sign across health states. In the absence of a chronic condition, smoking is found to have a tolerating effect (i.e., lower utility from a larger smoking history).

Several interesting trends emerge from these results. First, note that baseline marginal utility of both light and heavy consumption is negative with the exception of heavy smoking *with* a chronic condition. This finding is driven by the absence of an outside consumption good. Here, the only tradeoff from current smoking is the potential for future health consequences. As suggested by the rational addiction literature, the model cannot explain why individuals start smoking. The estimated preference parameters in the absence of a chronic illness suggest that, for a never smoker under the age of 25, an exogenous shock preference shock is required to get an individual to start smoking. That is, for the never smoker, the model predicts that smoking is unappealing. In addition to the negative marginal utility in the absence of a past smoking history, the dynamic considerations of the model suggest that smoking will increase the probability of future chronic illness and death through the smoking stock, which increases future smoking probabilities through the positive reinforcement effect, as well as, increasing the health marker index.

The second main trend from the estimated preference parameters is the reversal in sign of several preference parameters upon succumbing to a chronic illness. The baseline marginal utility of heavy smoking when in the chronic health state (α_{51}) flips to positive. Along with the positive reinforcement ($\alpha_{21}, \alpha_{61} > 0$) and the flip in the sign of the effect of the stock on utility ($\alpha_{101} > 0$), individuals now face a positive marginal utility from heavy smoking.

The model finds evidence of a small degree of individual variation in the effect of the smoking stock (A_{it}) on the health marker index (R_{it}) as the estimated standard deviation of θ, σ_θ, is nonzero. Recall further that the null hypothesis of σ_θ equaling zero is my explicit test for the presence of learning. While the results do suggest the presence of learning, the signals received at each health exam are quite noisy. The estimated standard deviation of the random error term (σ_ν) is large relative to $\overline{\theta}$ and σ_θ.

Table 6.1: Main Parameter Estimates

Description	Chronic Condition	Parameter	Estimate	ASE
Utility Parameters				
Constants				
	No	α_{00}	25.947	1.808
	Yes	α_{01}	1.364	0.272
Consumption - Light Smoking				
Constant	No	α_{10}	-6.128	0.117
Consumption*Smoking Stock	No	α_{20}	0.001	0.000
Consumption*Health Marker Index	No	α_{30}	0.000	0.000
Consumption*Age	No	α_{40}	0.070	0.001
Consumption	Yes	α_{11}	-7.479	0.194
Consumption*Smoking Stock	Yes	α_{21}	2.479	0.018
Consumption*Health Marker Index	Yes	α_{31}	-0.005	0.001
Consumption*Age	Yes	α_{41}	0.002	0.001
Consumption - Heavy Smoking				
Constant	No	α_{50}	-18.753	0.043
Consumption*Smoking Stock	No	α_{60}	1.704	0.010
Consumption*Health Marker Index	No	α_{70}	-0.001	0.000
Consumption*Age	No	α_{80}	0.000	0.000
Consumption	Yes	α_{51}	0.015	0.004
Consumption*Smoking Stock	Yes	α_{61}	2.483	0.018
Consumption*Health Marker Index	Yes	α_{71}	0.001	0.000
Consumption*Age	Yes	α_{81}	-0.068	0.004
Withdrawal				
	No	α_{90}	-6.927	0.046
	Yes	α_{91}	-1.539	0.133
Smoking Stock				
	No	α_{100}	-0.025	0.007
	Yes	α_{101}	2.636	0.051
Smoking Stock Squared				
	No	α_{110}	-0.002	0.001
	Yes	α_{111}	-0.596	0.005
Learning Parameters				
Mean Effect		$\overline{\theta}$	0.003	0.000
Standard Deviation of θ_i		σ_θ	0.098	0.004
Standard Deviation of ν		σ_ν	1.024	0.004
Additional Health Marker Index Parameters				
Lagged Health Marker Index		ζ	0.807	0.001
Age in Years		ϕ_1	0.005	0.000
Female		ϕ_2	-0.122	0.003
Education in Years		ϕ_3	-0.011	0.001
Married		ϕ_4	0.000	0.000
Constant		ϕ_5	1.039	0.013

Table 6.2 provides estimates of all other estimated model parameters. These estimates are not marginal effects and therefore are difficult to interpret because each outcome (health marker index, chronic health, death, etc.) is a complex function of entering period states and per-period decisions. In the simulation section below, I describe the results of simulations that isolate the effects of each variable on the system. However, a casual interpretation of

the results in Table 6.2 does yield some interesting insights. The parameter estimates of the smoking stock evolution equation indicate that an individual's stock of smoking depreciates faster than suggested by the medical literature. δ_1 suggests that, given cessation from smoking over the cycle of one year, the smoking stock is reduced by approximately 57%.[1] In the context of the model, 57% depreciation implies that after about six years of smoking cessation, an individual may have roughly the same health marker index and chronic health and death tranistion probabilities as a lifelong nonsmoker, all else equal. Additionally, the estimated magnitude of investment return in the smoking stock is greater for heavy compared to light smoking ($\delta_2 < \delta_3$).

As noted above, the estimated mean effect of the smoking stock on the health markers is positive ($\overline{\theta} = 0.003$). A greater smoking history therefore implies a higher, and thus worse health marker index. According to Table 6.2, a higher health marker index implies a higher probability of chronic illness (through the positive sign on λ_1), albeit at a decreasing rate (λ_2¡0), and death (through the positive signs on $\omega_1, \omega_2, \omega_3, \omega_5$, and ω_6. Furthermore, given a chronic illness, the probability of death is lower during the 1980s ($\omega_7 < 0$) and 1990s ($\omega_8 < 0$) both relative to before 1980 to capture exogenous advances in medical technology over time.

[1] Note that while this suggests a large amount of depreciation, the factor loading on unobserved heterogeneity for the stock equation slows that depreciation.

Table 6.2: Other Parameter Estimates

Description	Parameter	Estimate	ASE
Smoking Stock Parameters			
Depreciation Rate	δ_1	0.430	0.002
Investment, Light Smoking	δ_2	0.335	0.001
Investment, Heavy Smoking	δ_2	0.411	0.002
Standard Deviation of η	σ_η	0.134	0.000
Chronic Health Parameters			
Constant	λ_0	-12.040	0.057
Health Marker Index	λ_1	0.207	0.016
Health Marker Index Squared	λ_2	-0.010	0.001
1980s*Health Marker Index	λ_3	0.000	0.000
1990s*Health Marker Index	λ_4	0.000	0.000
Choice	λ_5	0.336	0.024
Choice*Health Marker Index	λ_6	-0.008	0.002
Age	λ_7	0.119	0.001
Education	λ_8	0.007	0.001
Gender	λ_9	0.019	0.005
Married	λ_{10}	-0.070	0.005
Mortality Parameters			
Constant	ω_0	-8.805	0.104
Health Marker Index	ω_1	0.001	0.000
Health Marker Index Squared	ω_2	0.001	0.000
Chronic Health State	ω_3	4.868	0.094
Choice	ω_4	0.000	0.000
Choice*Health Marker Index	ω_5	0.013	0.002
Choice*Chronic Health State	ω_6	0.503	0.021
1980s* Chronic Health State	ω_7	-0.086	0.017
1990s* Chronic Health State	ω_8	-0.214	0.030
Age	ω_9	0.041	0.002
Gender	ω_{10}	-0.061	0.014
Education	ω_{11}	-0.135	0.005
Married	ω_{12}	-0.204	0.028
Heterogeneity Parameters			
Utility: No Chronic Condition			
Not Smoking	ρ_{u00}	0.066	0.018
Light Smoking	ρ_{u01}	2.664	0.161
Heavy Smoking	ρ_{u02}	8.619	0.103
Utility: Chronic Condition			
Not Smoking	ρ_{u10}	0.964	0.148
Light Smoking	ρ_{u11}	-0.081	0.021
Heavy Smoking	ρ_{u12}	0.132	0.033
Stock	ρ_A	0.647	0.002
Health Marker Index	ρ_R	0.000	0.000
Chronic Health	ρ_H	0.001	0.000
Mortality	ρ_M	1.027	0.122
Mass Points and Probabilities			
Mass Point 1	μ_1	0.000	-
Mass Point 2	μ_2	1.270	0.122
Mass Point 3	μ_2	1.000	-
Coef. Weight on Mass Point 1	θ_1	-2.622	0.688
Coef. Weight on Mass Point 2	θ_2	-1.231	0.204
Miscellaneous Parameters			
Discount Factor	β	0.950	-
Log-Likelihood Value	$L(\Theta)$	-30481.266	

Mass points 1 and 3 are fixed at 0 and 1 respectively. Mass point 2 is estimated and its location is $\frac{exp(1.270)}{1+exp(1.270)} = 0.781$. The corresponding probabilities of mass points 1 through 3 are 0.053, 0.214, and 0.733.

The model is estimated with three points of support for the discretized unobserved heterogeneity distribution. The combination of positive estimated factor loadings on the marginal

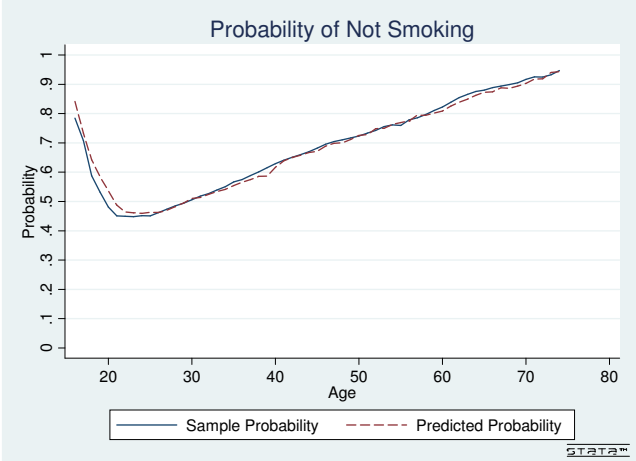
utility of smoking and on mortality provide evidence of positive selection. That is, those that are more likely to smoke are also more likely to die independently of smoking. These results suggest that there are certain permanent factors (e.g., propensity towards a risky lifestyle, genetics, etc.) that are correlated with both smoking and mortality. Interestingly, these unobserved factors are relatively uncorrelated with health markers or chronic health outcomes. However, the unobserved factors may still predict chronic disease induced mortality for higher type individuals if these factors drive smoking through the positive factor loadings on the marginal utility of smoking. By smoking more, these individuals build a larger smoking stock, and consequently are more likely to continue smoking through reinforcement. Because smoking predicts the onset of chronic illness, the unobserved factors may be indirectly linked to chronic illness caused mortality. Finally, because the model predicts that individuals only start smoking because of an exogenous shock to ϵ, for individuals of type three, smoking initiation may occur for a "smaller" shock because the positive factor loading on light and heavy smoking "offsets" some of the negative values for the direct marginal utility of smoking, $\alpha_{1.}$ and $\alpha_{5.}$.

6.2 Model Fit

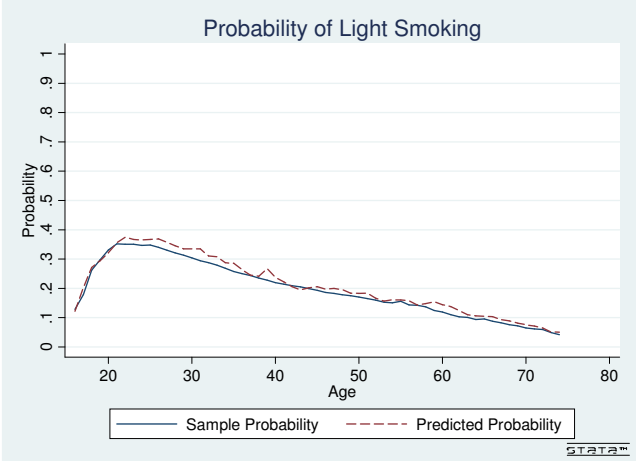
Figure 6.1 summarizes the relationship between the model's predicted probabilities and the observed data by age. Each pane of the figure represents one specific smoking option. For each individual, I compare observed smoking decisions and predicted smoking probabilities for periods up to either her final exam (exam seven) or death. I then average the results across individuals at each age.[2] The model predictions generated from the solution routine fit the data well even at ages for which there are not many observations.

Table 6.3 reports sample and predicted smoking probabilities by health exam and health state. I do not include a table on model fit by exam conditional on being in the chronic health state because less than one, three, and seven percent of individuals have a chronic condition in exams one, two, and three respectively. Note however that the average predicted choice probability across all health exams conditional on being in the chronic health state mirrors

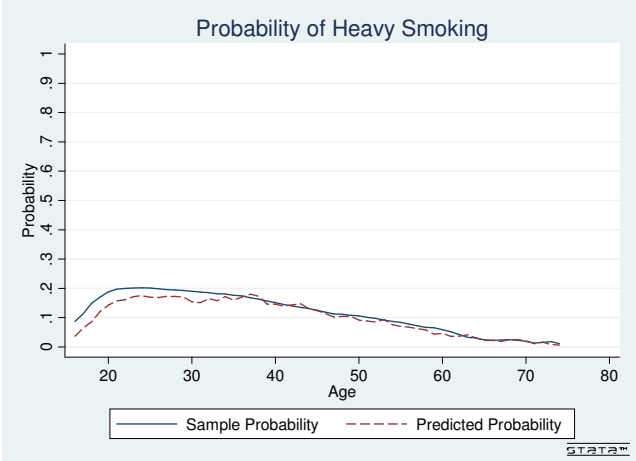
[2]Despite the fact that the model is solved from age 7 to 100, the figure only presents results for ages 20 to 75. Outside of the 20 to 75 age range, there are insufficient data for an informative comparison.



a.



b.



c.

Figure 6.1: Smoking Behavior by Age: Predicted and Sample Probabilities

the observed probabilities in the data fairly well. Table 6.3 suggests that the model does a good job of predicting whether or not an individual smokes at all. The model slightly under predicts light smoking and slightly over predicts heavy smoking.

Table 6.3: Model Fit: Choice Probabilities

	Not Smoking		**Light Smoking**		**Heavy Smoking**	
Exam	**Model**	**Data**	**Model**	**Data**	**Model**	**Data**
Unconditional on H_{it}. # Person/Year Obs.=19,461						
1	60.68	59.01	25.68	26.70	13.64	14.30
2	68.74	61.35	19.97	24.44	11.29	14.21
3	71.32	77.22	18.64	14.25	10.03	8.53
4	78.35	81.22	14.50	12.77	7.15	6.01
5	82.53	85.16	12.00	10.96	5.47	3.88
6	86.72	87.87	9.30	9.53	3.98	2.60
7	90.21	88.89	6.82	8.65	2.97	2.46
Mean	76.94	77.25	15.27	15.33	7.79	7.43
Conditional on $H_{it} = 0$. # Person/Year Obs.=17,601						
1	60.68	58.99	25.71	26.72	13.61	14.29
2	68.91	61.43	20.07	24.24	11.02	14.33
3	71.64	77.31	18.76	14.23	9.60	8.46
4	78.38	80.93	14.77	13.02	6.85	6.05
5	82.47	84.91	12.45	11.10	5.08	3.99
6	86.58	87.42	9.80	9.85	3.61	2.73
7	90.01	88.22	7.25	9.14	2.75	2.64
Mean	76.95	77.03	15.54	15.47	7.50	7.50
Conditional on $H_{it} = 1$. # Person/Year Obs.=1,860						
Mean	75.69	79.27	11.03	13.64	13.28	7.09

Figure 6.2 compares the observed sample probabilities of chronic health with the predicted health probabilities, as generated by the model at the estimated parameter values. As in Figure 6.1, Figure 6.2 averages predicted and sample probabilities across individuals by age only for those individuals with an observation at that age. Figure 6.2 reflects both transitions to and surviving members of the chronic health state. This is because solution to the model yields a

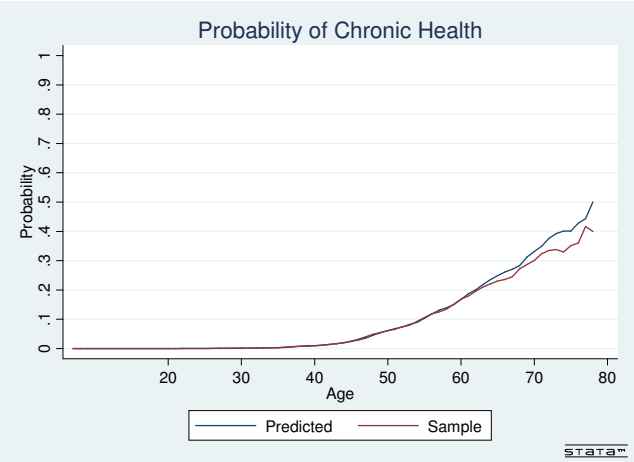


Figure 6.2: Chronic Health State by Age: Predicted and Sample Probabilities

predicted probability of transiting to a chronic health state of one for individuals already in that state. Note that for most ages, the model slightly over predicts the probability of being in a chronic health state.

Chapter 7

Simulation Design

In this chapter, I describe my simulation of smoking behavior and health outcomes using the structural model and the estimated parameters. Using my simulated sample, I address how learning from personalized information may impact smoking behaviors and health outcomes, as well as, how smoking affects morbidity and mortality outcomes. These results are presented in chapters 8 and 9, respectively. My simulations proceed as follows. First, I construct a simulated sample of 1000 individuals that mirrors the joint distribution of observable demographic characteristics (education, gender, marriage, and initial age upon entering the Framingham survey) of the Framingham sample. Next, for each simulated individual i, I construct 50 sets of match value, unobserved heterogeneity, and error draws over the estimated time frame.

$$\left\{\theta_{ik}, \mu_{ik}, \left\{\nu_{ikt}, \eta_{ikt}, \{\epsilon_{iktd}\}_{d=0}^{2}\right\}_{t=7}^{100}\right\}_{k=1}^{50}.$$

Smoking behavior and health outcomes are then simulated for each of the 50,000 observations from age seven until death.

First, I reconstruct Table 5.3 using the simulated smoking behavior to evaluate the model's performance in capturing overall smoking transitions and transitions around significant events. These results are reported in Table 7.1.[1] For those simulated to be not smoking in any given period, one period prior to a health exam, or one period prior to a chronic health shock, the

[1]Transitions around health exams are unconditional on chronic illness.

simulated smoking probabilities one and three periods after these events mirror those from the data. The model does less well in simulating behavior conditional on lagged light or heavy smoking. While the simulated probabilities of not changing behavior after one of the three events reflect those from the data, the model tends to under predict the probability of quitting and over predict the probability of switching to a different smoking intensity. However, the model does capture the general trend that more individuals have quit three years after an event when compared to one year after.

Table 7.1: Predicted Transitions

	Behavior One Period Post			Behavior Three Periods Post		
Behavior One Period Prior	Not Smoking	Light Smoking	Heavy Smoking	Not Smoking	Light Smoking	Heavy Smoking
Overall Transitions						
Not Smoking	97.80%	2.07%	0.09%	95.89%	3.30%	0.68%
Light Smoking	18.65	53.56	27.77	26.49	42.68	30.79
Heavy Smoking	3.62	31.62	64.74	7.73	35.05	57.18
Transitions Around Health Exams						
Not Smoking	98.73%	1.22%	0.05%	97.75%	1.98%	0.27%
Light Smoking	11.76	63.81	24.43	21.75	54.01	24.24
Heavy Smoking	4.66	37.90	57.44	9.91	42.49	47.60
Transitions Around Chronic Health Shocks						
Not Smoking	99.16	0.50	0.34	98.92	0.65	0.43
Light Smoking	32.78	44.09	23.13	55.86	27.55	16.59
Heavy Smoking	11.49	34.65	53.86	24.22	27.97	47.81

Chapter 8

Assessing the Effect of Information on Smoking Behavior

In this chapter, I use the simulate the model as described in chapter 7 to evaluate counterfactual scenarios that alter either the timing or the frequency with which information is received. First, to demonstrate the speed at which individuals learn, Table 8.1 reports the change in the average posterior variance after each health exam of the baseline simulation. Note that after the first exam (i.e., the first signal of information) the posterior variance decreases by nearly 20%. By the seventh exam, the mean posterior variance has been decreased by 40%. In spite of the "honing in" on individuals' true match values, smoking behavior appears to only slightly be influenced by learning. As a natural benchmark, I compare the

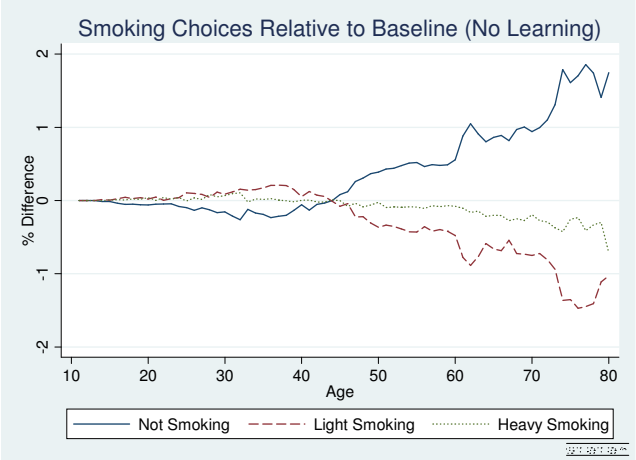
Table 8.1: Posterior Variance by Exam: The Speed of Learning

Exam	**Mean Posterior Variance**	**% Decrease**	**Cummulative % Decrease**
Initial Prior	0.0095	-	-
1	0.0076	19.9%	19.9%
2	0.0069	8.7%	26.9%
3	0.0066	4.7%	30.3%
4	0.0064	3.6%	32.8%
5	0.0061	4.0%	35.5%
6	0.0059	3.8%	37.9%
7	0.0057	2.9%	39.7%

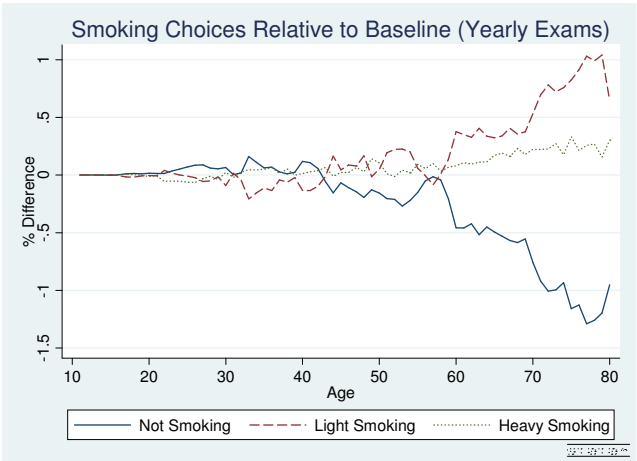
predictions of the baseline model to results from specifications with no learning (i.e., $\sigma_\theta = 0$), complete information (i.e., $\tau_{it} = \theta_i \ \forall t$), and a situation where an individual undergoes yearly

health exams as opposed to every four years. Figure 8.1 presents the mean percentage difference of simulated individuals choosing each smoking option for each information scenario relative to the baseline prediction. Somewhat counter intuitively, the simulations suggest that the effect of more information, that is, yearly exams, is only to encourage individuals to smoke lightly in later life. In the extreme, with complete information, individuals are more likely to smoke lightly at all ages. In both cases, there is no apparent change in heavy smoking. One possible explanation for this finding is that, because the effect of the smoking stock on the health marker index is small ($\overline{\theta} = 0.003$) and because the estimated standard deviation of the effect is large relative to the mean, upon learning their true match value, individuals feel that the health effects of smoking are manageable.[1] Ultimately, the effects of different information regimes are quit small. Even with yearly exams, by age 70, the difference in average smoking rates relative to the baseline model predictions are only approximately 0.06% higher.

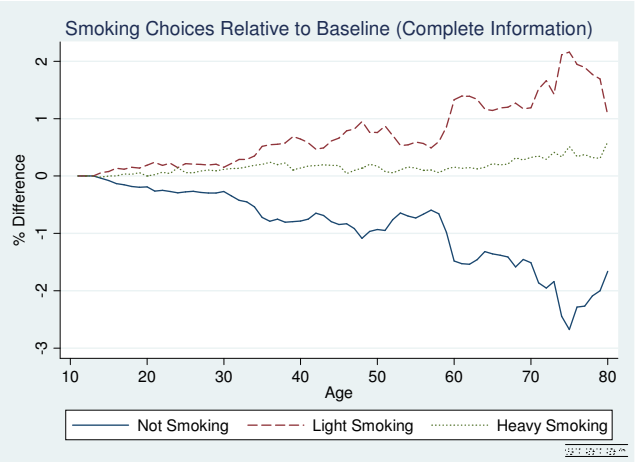
[1]For match values that are negative, there may be an incentive to continue to smoke because an increased smoking stock will decrease the health marker index, which in turn, will lower chronic health and death probabilities. Other experiments in which health signals where positively amplified, that is, while the health marker index evolved according to the estimated structural parameters, individuals *received signals* that suggested "scary" results, induced individuals to quit significantly more rapidly than the baseline results.



a.



b.



c.

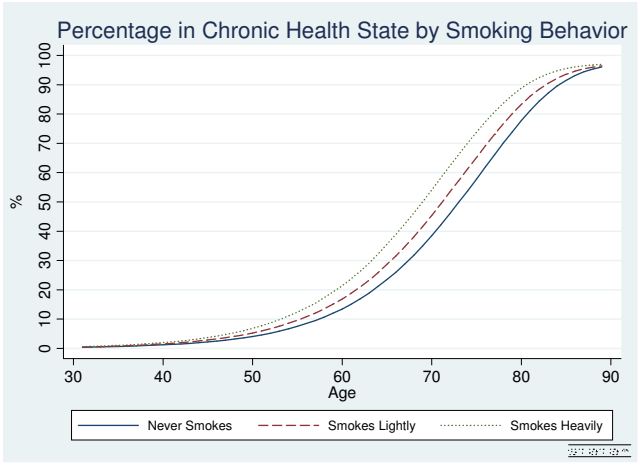
Figure 8.1: Average difference in smoking probabilities, relative to baseline choices, by age and across different policy scenarios

Chapter 9

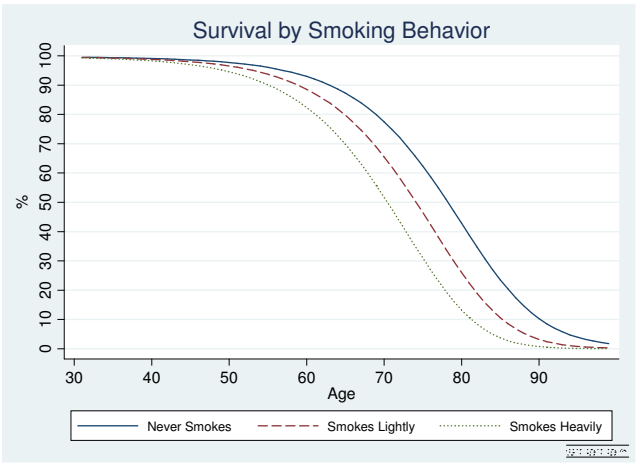
Assessing Health Effects and Health Selection

I use the simulated model to address how smoking impacts the age of chronic health onset and death. Figure 9.1a reports, by age, the percentage of the simulated sample with a chronic condition while forcing individuals to 1.) never smoke, 2.) smoke lightly from age 18, and 3.) smoke heavily from age 18.[1] Under these same forced behaviors, Figure 9.1b shows, by age, the percentage of the simulated sample that remains alive. The results in Figure 9.1 confirm the findings in Sloan *et al.* (2003) that the detrimental effects of smoking occur largely after the age of 50. Indeed, the gap in the percentage of the sample in the chronic health state between never smokers and heavy smokers widens from less than 10% at age 50 to more than 17% at age 70. Similarly, while the difference in those surviving to age 50 between heavy and never smokers is five percentage points, that gap widens to 30 percentage points at age 70. These results are roughly inline with those of Doll *et al.* (2004). Those authors find a difference of approximately 28 percentage points at age 70 when considering never smokers and smokers. The first half of Table 9.1 reports the mean age of onset for various health outcomes. Individuals who are forced to smoke lightly and smoke heavily from age 18 onwards face a mean age of chronic health onset that are approximately two and four years earlier than those forced to never smoke. While Doll *et al.* (2004) report that smoking shortens the lifespan

[1] Recall from the structural model that I assume that, upon transiting to a chronic health state, an individual remains in that state for life.



a.



b.

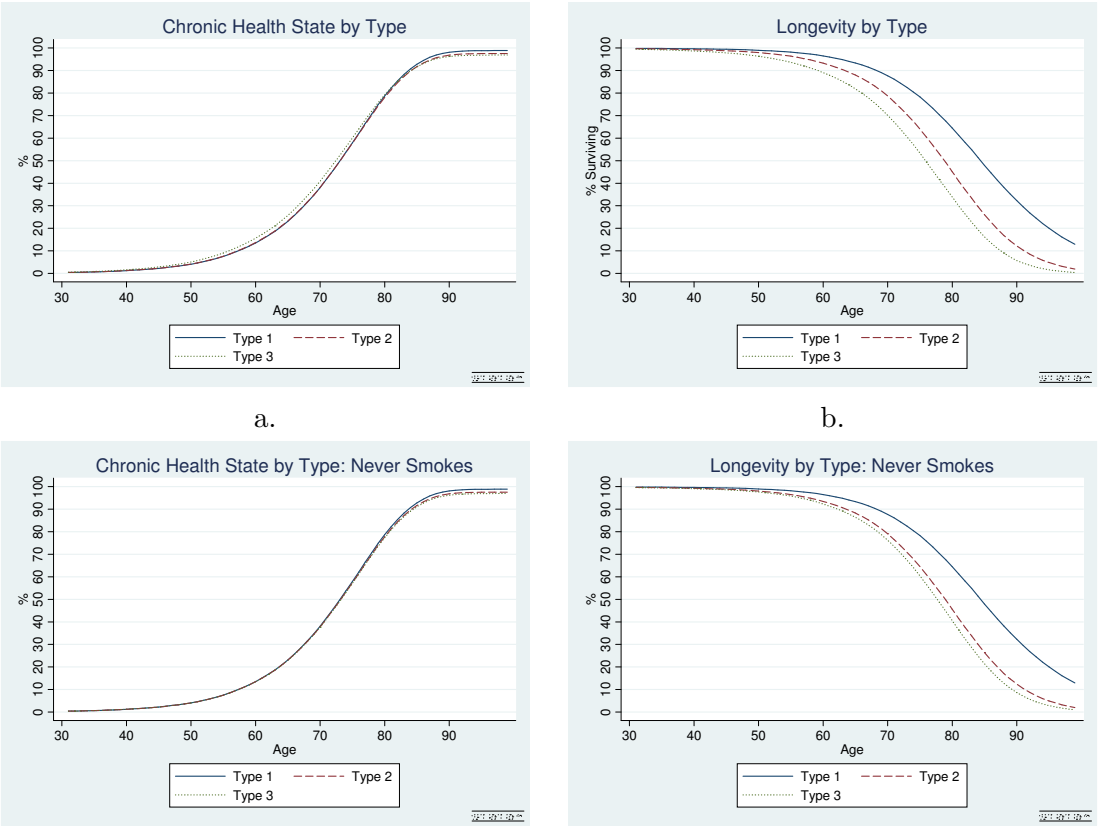
Figure 9.1: Percentage of simulated sample a.) in the chronic health state and b.) remaining, by age and quit status

by ten years, my results suggest the reduction is approximately four and eight.[2]

While Doll *et al.* (2004) only condition their results on decade of birth and gender, I report results that are conditional on both observed and unobserved factors. Here, I highlight the importance of incorporating unobserved heterogeneity. Figure 9.2 plots the same two graphs as in Figure 9.1 but now conditions each result by unobserved "type". Panels a. and b. report health outcomes under the baseline rational choice simulation whereas panels c. and d.

[2]Doll *et al.* (2004) do not take into account intensity of smoking in these calculations. My results indicate that, conditional on smoking, the intensity with which one smokes is an important factor explaining health outcomes.

report health outcomes assuming that all simulated individuals never smoke. Note that while unobserved heterogeneity does not play a significant role in chronic health transitions, the model predicts that type three individuals face lower expected longevity in both the baseline and nonsmoking simulations. Recall that the alternative specific factor loadings in the utility function greatly increase the marginal utility of smoking for individuals of a higher type. Indeed, the model predicts that only individuals with the largest mass point, type three, will ever smoke. Therefore, Figure 9.2 demonstrates that, independent of smoking, individuals of a higher type face lower expected longevity.



a. b.

c. d.

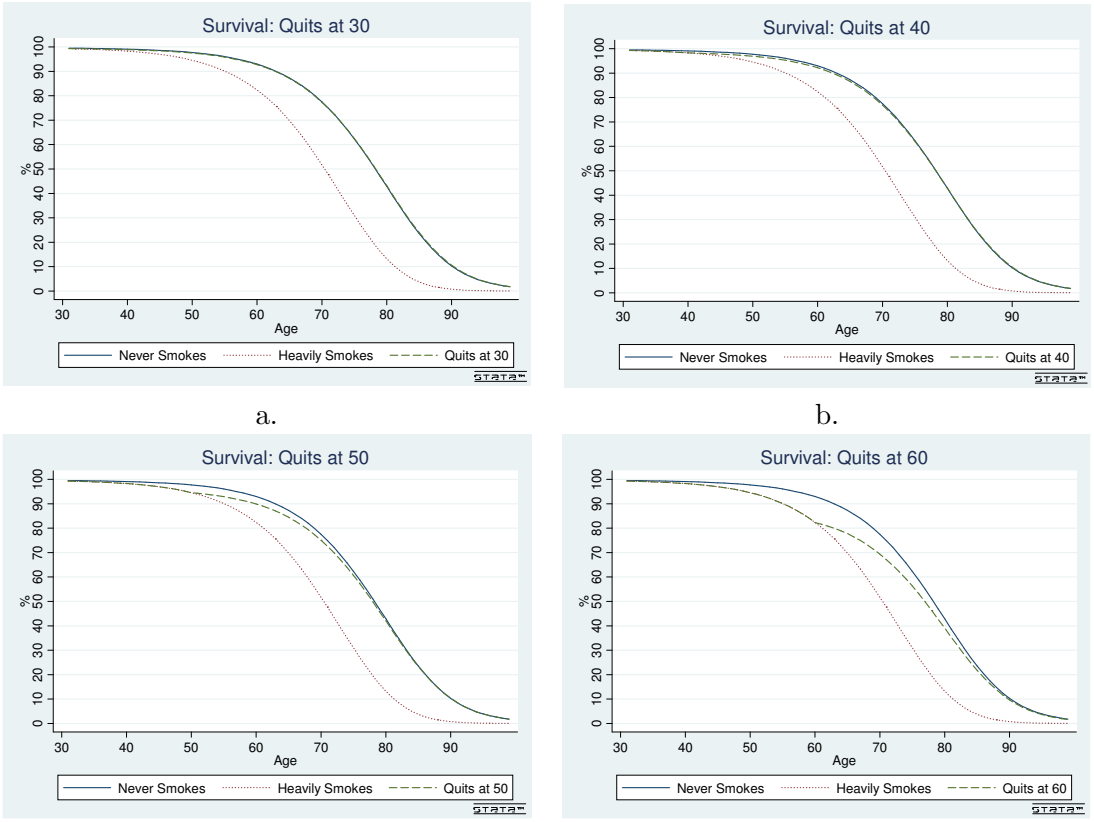
Figure 9.2: a. Simulated chronic health state by age at baseline. b. Simulated longevity by age at baseline. c. Simulated chronic health state by age assuming no smoking. d. Simulated longevity by age assuming no smoking.

Next, I use the model to simulate chronic health and death outcomes under different lifetime smoking paths to assess the impact of smoking cessation on these outcomes. I simulate health outcomes assuming that an individual smokes heavily from age 18 and quits forever at ages 30, 40, 50, and 60. The results, reported in Table 9.1, imply that quitting smoking at ages 30, 40, 50, and 60 years of age increases life-expectancy by approximately 8, 7.75, 7, and 5.5 years, respectively. These results suggest clear life expectancy gains from quitting at *all stages of the life cycle*.

Table 9.1: Age of Chronic Health Onset and Death

Variable	**Mean Age of Chronic Health Onset**	**Mean Age of Death**
Never Smokes	70.75	77.60
	(10.72)	(11.60)
Smokes $\leq$ 1 Pack/day from Age 18	68.91	73.32
	(10.89)	(11.19)
Smokes $>$ 1 Pack/day from Age 18	66.79	69.58
	(10.87)	(10.94)
Smokes $>$ 1 Pack/day from Age 18 *and*		
quits at Age 30	70.77	77.58
	(11.00)	(11.85)
quits at Age 40	70.54	77.32
	(11.33)	(12.28)
quits at Age 50	69.98	76.60
	(11.83)	(11.02)
quits at Age 60	66.55	74.99
	(10.91)	(13.85)

Figure 9.3 shows the survival percentages by age for the different smoking patterns. Note that for individuals that quit at age 30, their expected longevity is roughly identical to never smokers. Similarly, quitting by age 40 has minimal effects on mortality probabilities. Individuals that smoke into their fifties and sixties, however, have a much more likely chance of dying prematurely.



a. b. c. d.

Figure 9.3: Percentage of simulated sample remaining, by age and quit status

Chapter 10

Conclusion

This study formulates and estimates a dynamic stochastic model of smoking behavior. The model extends the classic rational addiction model to allow for health learning. By estimating the structural parameters of the model, I capture preferences and expectations in the tradeoff between smoking and the potential for future health shocks. The structural approach also allows for counterfactual simulations that a.) assess the importance of health marker information in the decision to smoke cigarettes, and b.) capture the direct effect of smoking, and smoking cessation, on different health outcomes while controlling for unobserved heterogeneity.

Generally, I find that significant reinforcement and withdrawal effects drive smoking dynamics by altering the future marginal utilities of smoking. The reinforcement effect is estimated with a novel construction of past smoking behavior that is consistent with the theoretical notion of a "smoking stock". I empirically construct this stock as a weighted average of several measures of smoking history using using principal components analysis. Consistent with the theory, individuals understand that different smoking choices influence the smoking stock through depreciation and investment coefficients. The smoking stock then reinforces future smoking, through the marginal utility of smoking, and influences health, through the health marker index.

Estimates of the structural parameters suggest that there exists heterogeneity in the effect of the accumulated smoking stock on an index of health markers. Despite this heterogeneity, learning about an individual's own place in the distribution of this effect, at least in the Framingham Heart Study setting, does not appear to significantly inform smokers about the

long-term health consequences of smoking. In fact, learning about how smoking effects health markers may actually slightly increase moderate smoking in older individuals. However, individuals that receive sharp, discrete shocks that imply worse health markers typically do scale down their smoking behavior (either by quitting or lowering the intensity with which they smoke). Only when a health marker shock pushes an individual's health markers into the worst 10% does the resulting change in beliefs have a significant effect on smoking behavior. Therefore, my results are consistent with the literature on personalized health information that have found changes in smoking behavior after serious health shocks (Smith *et al.*, 2001; Sloan *et al.*, 2003; Khwaja *et al.*, 2006; Arcidiacono *et al.*, 2007).

The lack of a change in smoking to marginal changes in health markers, observed from Framingham Heart Study health exams, may be evidence of some limitations of this study. First, while the FHS administers health exams every two to eight years, individuals may be observing their health markers at other doctor visits. Interim doctor visits may induce individuals to quit smoking; however, the econometrician would only observe that the individual quit. With respect to the model, the observed quitting would be attributed to the preference error, ϵ. In this case, the conclusion would not be that health markers are uninformative as to the implications of smoking, but rather that the Framingham Heart Study signals of information do not provide any additional information to what is already known. Because of this possibility, I interpret my results on as a lower bound on health marker learning.

Second, as noted in Sloan *et al.* (2003), smoking behavior may be altered by a change in risk perceptions, which may be changed by new information. Because of the absence of subjective expectation data regarding the effect of smoking on health markers, I am required to place more structure on the learning process. Thus, the assumption of conjugate normal distributions for the signal of information and beliefs may be driving the results. Alternatively, I could specify a beta/binomial learning process. Here, a discrete signal of information (e.g., blood pressure exceeding some threshold) causes individuals to update their continuous belief distribution.[1] However, learning about multiple health marker discrete shocks would

[1]Recall that a central goal of this dissertation is to examine the effects of health information *prior* to major health events. While a heart attack, say, is clearly a discrete health shock that may greatly alter risk perceptions

be computationally intractable. Furthermore, it is not clear how one could both define the relevant signal (i.e., which health marker) and it's corresponding threshold. I leave different specifications of the learning process for future work.

I find evidence of positive selection with respect to smoking and mortality by estimating the correlation in permanent unobserved heterogeneity between these outcomes. Factor loadings that dictate the effect of the permanent unobserved term on the marginal utility of smoking and on mortality are both estimated to be positive. This finding suggests that individuals that are more likely to smoke, are also more likely to die *independently of smoking*. Interestingly, the factor loadings that capture the correlation in permanent unobservables across smoking and chronic health and health markers are not statistically different from zero. This finding suggests that, while there exists an unobserved relationship between the propensity to smoking and mortality, the excess in mortality cannot be attributed directly to chronic disease (as defined in this paper) or health markers. However, as noted above, if unobserved factors drive certain individuals to smoke, and smoking predicts the onset of chronic conditions, the unobserved factors may still predict disease related mortality through smoking.

Simulations of the structural model confirm the positive selection and suggest that, when controlling for unobserved heterogeneity, the effects of smoking on mortality outcomes may be less extreme than previously estimated. I find that smoking heavily from age 18 can reduce life expectancy by eight years relative to life-long non-smokers and by four years relative to those smoking only lightly ($\leq$ 1 pack/day) from age 18. I compare my results to those of Taylor *et al.* (2002); Doll *et al.* (2004); Brønnum-Hansen *et al.* (2007) that find overall longevity loss from daily smoking to be roughly 7.4-10.5, 10, and 8.7-10.4 years, respectively. Furthermore, quitting smoking by age 30 implies relatively few chronic health or mortality differences, on average, from life-long non-smokers; however, waiting to quit until age 60 implies that the health consequences may be severe. Indeed, as suggested by the literature, the major effects of smoking on health are realized after age 50 (Sloan *et al.*, 2003). With respect to the estimated health effects, an important limitation of this dissertation is the definition of a

and smoking behavior, the focus here is on health markers.

chronic condition. Due to data limitations, the dichotomous variable for chronic health used throughout does not capture all diseases that are caused by smoking. For example, the Framingham Heart Study data do not include panel data for chronic obstructive pulmonary disease (COPD). Given that COPD is the number four leading cause of death in the United States[2], the omission of COPD in the chronic health indicator may understate the importance of both disease in the probability of death and the extent to which preferences for smoking vary across chronic health states. Furthermore, I can only assess the correlation in permanent unobserved factors across smoking, mortality, and cardiovascular and cancer related chronic health.

The two main questions of this dissertation should guide future work. First, are there sources of information, personalized or otherwise, that effectively convince individuals to stop smoking that also are not major health shocks? Do health markers influence smoking behavior in settings other than the Framingham Heart Study? It would be interesting to collect subjective expectation data on risk perceptions in which surveyors explicitly mention individual specific variation in health markers. Would these data show a role for health markers to change risk perceptions and subsequent smoking behavior? Second, what are the sources of unobserved heterogeneity that are shown to be correlated across preferences for smoking and mortality. Would controlling for and modeling parental smoking behavior or parental health outcomes significantly change the role of unobserved heterogeneity in the results of this dissertation? Furthermore, how would the unobserved heterogeneity distribution change if alcohol consumption was explicitly modeled? In future work, I hope to answer these questions.

[2] http://www.nhlbi.nih.gov/health/dci/Diseases/Copd/Copd_WhatIs.html

Appendix A

Bayesian Updating

Here, I derive the posterior beliefs discussed in the main text (Equations 4 and 5). I assume rational expectations such that an individual's initial belief upon entering the sample regarding their true θ_i is the population distribution:

$$E_0(\theta_i) = \tau_{i0} = \overline{\theta}$$

$$V_0(\theta_i) = \psi_{i0} = \sigma_\theta^2.$$

Consider an individual in period t with smoking stock A_{it}. For ease of exposition, assume that an individual takes a health exam each period. When deriving the posterior beliefs in period t, an individual considers only her prior beliefs $(\tau_{it-1}, \psi_{it-1})$ and her signal of information k_{it}. According to Bayes' Rule, the posterior distribution, f_t, of θ_i is given as:

$$f_t(\theta_i|\kappa_n, \tau_{it-1}, \psi_{it-1}) \propto f_{t-1}(\theta_i) g(\kappa_{it}|A_{it}, \theta_i, \sigma_\nu). \tag{A.1}$$

Note that while $g(\kappa_{it}|A_{it}, \theta_i, \sigma_\nu)$ conveys information about κ^{it}, an individual knows A_{it} and, because θ_i is time invariant, can therefore infer information about θ_i over time. This will become more clear in the interpretation of the posterior mean and variance. First consider $g(\kappa_{it}|A_{it}, \theta_i, \sigma_\nu)$:

$$g(\kappa_{it}|A_{it}, \theta_i, \sigma_\nu) = \frac{1}{(2\pi\sigma_\nu^2)^{\frac{1}{2}}} \exp\Big(\frac{1}{2\sigma_\nu^2}(\kappa_{it} - \theta_i A_{it})^2\Big). \tag{A.2}$$

Note that because we are concerned with the distribution of θ_i, any term that does not include θ_i can be treated as part of the normalizing constant. We can ignore the first term within the

parenthesis:

$$\propto \exp\Big(\frac{-1}{2\sigma_\nu^2}\big(-2\theta_i\kappa_{it}A_{it}+\theta_i^2A_{it}^2\big)\Big).$$

Simplifying and completing the square yields:

$$\propto \exp\Big(-\frac{A_{it}^2}{2\sigma_\nu^2}\big(\theta_i-\frac{\kappa_{it}A_{it}}{A_{it}^2}\big)^2\Big).$$

Notice that the term subtracted from θ_i is the within (individual i) variation ordinary least squares estimate of θ_i from the n^{th} signal of information. Define $\hat{\theta_{it}} = \frac{\kappa_{it}A_{it}}{A_{it}^2}$. Substituting for $\hat{\theta_{it}}$, we have that:

$$g(\kappa_{it}|A_{it},\theta_i,\sigma_\nu) \propto \exp\Big(-\frac{A_{it}^2}{2\sigma_\nu^2}\big(\theta_i-\hat{\theta_{it}}\big)^2\Big). \tag{A.3}$$

Now consider the prior probability distribution of θ_i:

$$f_{t-1}(\theta_i) = \frac{1}{(2\pi\psi_{it-1})^{\frac{1}{2}}}\exp\Big(\frac{1}{2\psi_{it-1}}(\theta_i-\tau_{it-1})^2\Big). \tag{A.4}$$

The nice aspect of the conjugate distribution assumption is that we can characterize the posterior distribution sufficiently with closed form expressions for the posterior mean and variance. Therefore, we only have to characterize that part of the posterior density that captures the mean and variance. In that light, consider the product of the exponential portions of Equations A.3 and A.4 after rearranging terms and absorbing those without θ_i into the normalizing constant:

$$f_t(\theta_i) \propto \Bigg(-\frac{1}{2\psi_{it-1}\sigma_\nu^2}\Big(\theta_i^2\big(A_{it}^2\psi_{it-1}+\sigma_\nu^2\big)-2\theta_i\big(A_{it}^2\psi_{it-1}\hat{\theta}_{it}+\sigma_\nu^2\tau_{it-1}\big)\Big)\Bigg). \tag{A.5}$$

After rearranging and completing the square, we have the kernel of a normal distribution representing the posterior distribution:

$$f_t(\theta_i) \propto \Bigg(-\frac{A_{it}^2\psi_{it-1}+\sigma_\nu^2}{2\psi_{it-1}\sigma_\nu^2}\Big(\theta_i-\big(\frac{A_{it}^2\psi_{it-1}\hat{\theta}_{it}+\sigma_\nu^2\tau_{it-1}}{A_{it}^2\psi_{it-1}+\sigma_\nu^2}\big)^2\Big)\Bigg).$$

The posterior mean and variance is:

$$\tau_{it} = E(\theta_i|\kappa_t, \tau_{it-1}, \psi_{it-1}) = \left(\frac{A_{it}^2\psi_{it-1}}{A_{it}^2\psi_{it-1} + \sigma_\nu^2}\right)\hat{\theta}_{it} + \left(\frac{\sigma_\nu^2}{A_{it}^2\psi_{it-1} + \sigma_\nu^2}\right)\tau_{it-1} \tag{A.6}$$

$$\psi_t = Var(\theta_i|\psi_{it-1}, \sigma_\nu) = \frac{\psi_{it-1}\sigma_\nu^2}{A_{it}^2\psi_{it-1} + \sigma_\nu^2}. \tag{A.7}$$

Rearranging these equations yields the posterior mean and variance equations above.

Bibliography

Ackerberg, D. (2003). Advertising, Learning, and Consumer Choice in Experience Good Markets: An Empirical Examination. *International Economic Review*, **44** (3), 1007–1040.

Adda, J. and Lechene, V. (2001). Smoking and Endogenous Mortality: Does Heterogeneity in Life Expectancy Explain Differences in Smoking Behavior, discussion Paper Series, Department of Economics, University of Oxford, ISSN 1471-0498.

— and — (2004). On the identification of the effect of smoking on mortality, discussion Paper Series, Department of Economics, University of Oxford, Working Paper Number 184.

Aguirregabiria, V. and Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, **156** (1), 38–67.

Arcidiacono, P., Sieg, H. and Sloan, F. A. (2007). Living Rationally Under the Volcano? An Empirical Analysis of Heavy Drinking and Smoking. *International Economic Review*, **48** (1), 37–65.

Becker, G. S., Grossman, M. and Murphy, K. (1994). An Empirical Analysis of Cigarette Addiction. *American Economic Review*, **84** (3), 396–418.

— and Murphy, K. (1988). A Theory of Rational Addiction. *Journal of Political Economy*, **96** (4), 675–700.

Bennett, C. H. and Richardson, D. R. (1984). Effects of Chronic Tobacco Smoke Exposure on Arterial Blood Pressure Regulation. *American Journal of Physiology: Heart and Circulatory Physiology*, **247**.

Benowitz, N. L. (2003). Cigarette Smoking and Cardiovascular Disease: Pathophysiology and Implications for Treatment. *Progress in Cardiovascular Diseases*, **46** (1), 91–111.

Bernheim, B. D. and Rangel, A. (2004). Addiction and cue-triggered decision processes. *American Economic Review*, **94** (5), 1558–1590.

Blau, D. M. and Gilleskie, D. B. (2008). The Role of Reitree Health Insurance in the Employment Behavior of Older Men. *International Economic Review*, **49** (2), 475–514.

Brønnum-Hansen, H., Juel, K., Davidsen, M. and rensen, J. S. (2007). Impact of selected risk factors on expected lifetime without long-standing limiting illness in denmark. *Preventive Medicine*, **45**, 49–53.

Carbone, J. C., Kverndokk, S. and Rogeberg, O. J. (2005). Smoking, Health, Risk, and Perception. *Journal of Health Economics*, **24**, 631–653.

Chaloupka, F. (1991). Rational Addiction Behavior and Cigarette Smoking. *Journal of Political Economy*, **99** (4), 722–742.

Chan, T. Y. and Hamilton, B. (2006). Learning, Private Information, and the Economic Evaluation of Randomized Experiments. *Journal of Political Economy*, **114** (6), 997–1040.

Chernew, M., Gowrisankaran, G. and Scanlon, D. P. (2008). Learning and the Value of Information: Evidence from Health Plan Report Cards. *Journal of Econometrics*, **144**, 156–174.

Crawford, G. S. and Shum, M. (2005). Uncertainty and Learning in Pharmaceutical Demand. *Econometrica*, **73** (4), 1137–1173.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M. and Kannel, W. B. (2008). General Cardiovascular Risk Profile for Use in Primary Care. *Circulation*, **117** (6), 743–753.

Doll, R., Peto, R., Boreham, J., Gray, R. and Sutherland, I. (2004). Mortality in Relation to Smoking: 50 Years' Observations on Male British Doctors. *British Medical Journal*, **328**, 1519–1528.

—, —, Wheatley, K., Gray, R. and Sutherland, I. (1994). Mortality in Relation to Smoking: 40 Years' Observations on Male British Doctors. *British Medical Journal*, **309**, 901–911.

Epstein, L. G. and Zin, S. E. (1991). Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis. *The Journal of Political Economy*, **99** (2).

Garrison, R., Kannel, W., Feinleib, M., Castelli, W., McNamara, P. and Padgett, S. (1978). Cigarette Smoking and HDL Cholesterol. *Atherosclerosis*, **30**, 17–25.

Gilleskie, D. B. (1998). A Dynamic Stochastic Model of Medical Care Use and Work Absence. *Econometrica*, **66** (1), 1–45.

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, **80** (2), 223–255.

Gruber, J. and Koszegi, B. (2001). Is Addiction "Rational"? Theory and Evidence. *The Quarterly Journal of Economics*, **116** (4).

Gul, F. and Pesendorfer, W. (2007). Harmful addiction. *Review of Economic Studies*, **74**, 147–172.

Heckman, J. J. and Singer, B. (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, **52** (2), 271–320.

Jha, P., Peto, R., Zatonski, W., Boreham, J., Jarvis, M. J. and Lopez, A. D. (2006). Social inequalities in male mortality, and in male mortality from smokin: indirect estimation from national death rates in england and wales, poland, and north america. *Lancet*, **368**, 367–370.

Keane, M. P. and Wolpin, K. I. (1994). The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence. *The Review of Economics and Statistics*, **76** (4).

Khwaja, A. (2010). A Life Cycle Analysis of the Effects of Medicare on Individual Health Incentives and Health Outcomes. *Journal of Econometrics*, **156**, 130–147.

—, Sloan, F. A. and Chung, S. (2006). Learning About Individual Risk and the Decision to Smoke. *International Journal of Industrial Organization*, **24** (), 683–699.

Kreps, D. M. and Porteus, E. L. (1978). Temporal Resolution of Uncertainty and Dynamic Choice Theory. *Econometrica*, **46** (1).

— and — (1979). Dynamic Choice Theory and Dynamic Programming. *Econometrica*, **47** (1).

Mira, P. (2007). Uncertain Infant Mortality, Learning, and Life-Cycle Fertility. *International Economic Review*, **48** (3), 809–846.

Mroz, T. (1999). Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome. *Journal of Econometrics*, **92** (2), 233–274.

Omvik, P. (1996). How Smoking Affects Blood Pressure. *Blood Pressure*, **5**, 71–77.

Orphanides, A. and Zervos, D. (1995). Rational Addiction with Learning and Regret. *Journal of Political Economy*, **103** (4), 739–758.

Rust, J. (1987). Optimal Replacement of Gmc Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica*, **55** (5), 999–1033.

— and Phelan, C. (1997). How Social Security and Medicare Affect Retirement Behavior in a World With Incomplete Markets. *Econometrica*, **65** (4), 781–832.

Sickles, R. C. and Williams, J. (2008). Turning from Crime: A Dynamic Perspective. *Journal of Econometrics*, **145**, 158–173.

Sloan, F. A., Smith, V. K. and Taylor, D. H. (2003). *The Smoking Puzzle: Information, Risk Perception, and Choice*. Harvard University Press.

Smith, V. K., Taylor, D. H., Sloan, F. A., Johnson, F. R. and Desvousges, W. H. (2001). Do Smokers Respond to Health Shocks? *The Review of Economics and Statistics*, **83** (4), 675–687.

Suranovic, S. M., Goldarb, R. S. and Leonard, T. C. (1999). An Economic Theory of Cigarette Addiction. *Journal of Health Economics*, **18**, 1–29.

Taylor, D. H., Hasselblad, V., Henley, S. J., Thun, M. J. and Sloan, F. A. (2002). Benefits of Smoking Cessation for Longevity. *American Journal of Public Health*, **92** (6), 990–996.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edn.

United States Department of Health and Human Services (1990). The Health Benefits of Smoking Cessation: A Report of the Surgeon General's Report.

United States Department of Health and Human Services (2004). The Health Consequences of Smoking: A Report of the Surgeon General.

United States Department of Health and Human Services (2010). How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: Surgeon General's Report.

Viscusi, W. K. (1990). Do Smokers Underestimate Risks? *Journal of Political Economy*, **98** (6), 1253–1269.

— and Evans, W. N. (1990). Utility Functions that Depend on Health Status: Estimates and Economics Implications. *American Economic Review*, **80** (3), 353–374.

— and Hakes, J. K. (2008). Risk Beliefs and Smoking Behavior. *Economic Inquiry*, **46** (1), 45–59.

Vyas, S. and Kumaranayake, L. (2006). Constructing Socio-Economic Status Indices: How to Use Principal Component Analysis. *Oxford University Press in association with the London School of Hygiene and Tropical Medicine*.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H. and Kannel, W. B. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, **97**, 1837–1847.

Wray, L. A., Herzog, A. R., Willis, R. J. and Wallace, R. B. (1998). The Impact of Education and Heart Attack on Smoking Cessation amoung Middle-Aged Adults. *Journal of Health and Social Behavior*, **39**, 271–294.