Obesity and Health-Related Decisions: an Empirical Model of Weight Status for Young Adults in the US.

By Leonardo Fabio Morales[1]

Abstract

Obesity is widely accepted as one of the main causes of premature death, and the causal relationship between obesity and several of the most deadly chronic diseases is a consensus in the medical and public health literature. Obesity in the United States has recently been recognized as a public health concern and a social problem because the rise in the obesity prevalence rate has been stunning over the past three decades. Using AddHealth, a longitudinal study of teenagers and young adults in the United States, I estimate a comprehensive dynamic model of obesity determination that assumes as endogenous several factors mentioned in the literature as obesity determinants: physical activity, smoking, a proxy for food consumption, and childbearing. Two additional endogenous decisions included in the model are career-related decisions and residential location decisions. The first is included because it determines the intensity, in terms of energy expenditure, of individuals' daily main activities. The second is included because it determines the built environments in which individuals live. I specify reduced form equations for all these endogenous demand decisions, together with an obesity structural equation. The whole system of equations is jointly estimated by full information log-likelihood methods. The errors in all equations are assumed to be correlated with each other in the estimation. I use the discrete factor random effects estimation method to model this unobserved heterogeneity. Using the empirical model to study the mechanisms behind the determination of obesity, I am able to quantify the effect on the probability of obesity of several individual decisions after controlling for the endogenous nature of those decisions. This research provides evidence of important effects of physical activity on the reduction of the probability of obesity for young men and women. In addition, I found evidence of a small but significant negative effect of the availability of a set of neighborhood amenities on the probability of male and female obesity. This is an important contribution of this research to the literature, because these results are obtained from a framework in which the residential location of individuals is explicitly modeled as an endogenous decision. Up to this author knowledge, there have been no attempts to model residential location decisions in studies on obesity.

1 Introduction

In recent years economists, especially in the field of health economics, have shown an increased interest in health outcomes associated with the weight status of individuals. This growth of interest is not surprising because obesity has been strongly related in the medical and public health literature to chronic diseases

[1] This is a preliminary and incomplete work. All errors are entirely the author's responsibility.

such as diabetes type II, heart disease, and hypertension (Mokdad et al., 2001; Must et al., 1999). In addition, the prevalence of obesity has risen to such a degree in developed countries that it is now considered an epidemic. For the United States, in 2008 the prevalence rate of obesity was 32.2% among adult men and 35.5% among adult women (Flegal et al., 2010). These rates imply a dramatic increase in the last three decades when compared with the prevalence of 12.7% for men and 17% for women measured in the late 1970s (Eid et al., 2008).

There is an ongoing debate about the factors that cause obesity and that have contributed to the remarkably high obesity prevalence in the U.S. Several studies in the literature on obesity have focused on the effect that relative prices of calories and physical activity have on the determination of weight status. More in line with the purposes of this research are the efforts that have been made to find causal links between individual choices and obesity measures. The obesity causal factors usually taken into account in the literature are smoking, physical activity, diet, and similar individual lifestyle descriptors. Some of these papers have noted that individual choices usually associated with obesity are endogenous (Rashad, 2006; Ng et al., 2010). Another factor that has been explored in research about the determinants of obesity is the environment in which individuals perform their daily activities. Recent literature in epidemiology, urban economics, and planning has focused on the role of built environments in increasing energy consumption and decreasing energy expenditure (Papas, 2007). In other words, neighborhoods may affect demand for exercise and diet, and thereby have an impact on obesity.

In this dissertation I propose a theoretical and empirical framework for modeling weight status and additional endogenous individual behaviors that may play important roles in the determination of an individual's weight. Within this framework, the probability of being obese is the result of endogenous choices, exogenous factors, and an unobserved heterogeneity component. Econometrically, the estimation strategy used here consists of the specification of a system of equations that include weight status and the set of endogenous choices. The entire system is jointly estimated by full information maximum likelihood methods. The estimation technique also incorporates unobserved heterogeneity in the equations by using a semi-parametric method that does not require assumptions about its distribution.

In addition to taking into account lifestyle choices (smoking, physical activity, etc.) that have been linked to obesity in the literature, this research also incorporates two major decisions in individuals' lives: career-related decisions and residential-location decisions. Inclusion of these choices is an important extension of obesity models for two reasons. One is that individual energy expenditure levels depend greatly upon the kind of career path a person decides to follow, because there are different levels of physical activity for different jobs, different resources for healthier lifestyles in different professions, etc. The second is that residence location decisions determine the characteristics and resources of neighborhoods in which individuals live. These characteristics and resources may encourage individuals to increase their energy expenditure levels by engaging in physical activity. By modeling the residential

location decision, I am able to control for the potential endogeneity of neighborhood characteristics in the decision to perform any sort of physical activity. This is one of the most important contributions of this research to the literature. Modelling residential decisions is crucial because the effect of neighborhood characteristics on obesity will be biased if researchers ignore the fact that individuals self-select themselves into their neighborhoods. This research is a step forward in this direction because, up to this author knowledge, there have been no attempts to model residential location decisions in studies on obesity.

Using the estimated model, I measure the contribution of several endogenous factors to the probability of an individual being obese. Another special feature of this model is that it allows me to test the hypothesis that different neighborhood amenities have different impacts on individuals' endogenous lifestyle decisions, such as the performance of physical activity. Therefore, this research may contribute to the recent debate about the influence of built environments on the propensity to become obese. Several findings of a causal relationship with regard to this hypothesis have been criticized for assuming that environmental factors are exogenous. Critics are motivated by the facts that 1) the environment is usually represented by neighborhood characteristics and 2) these characteristics are endogenous, because residential selection is an individual's choice.

Although the existence of an obesity epidemic in the United States is well established, the kind of public policy that would be effective in dealing with the problem remains unclear. The present research contributes to this debate by proposing a comprehensive model of obesity determination. This model allows exploration of the contribution of several endogenous decisions to the probability of being obese in a framework that controls for the endogeneity of these choices. With the estimated model I perform some experiments that allow us to see what the evolution of the obesity prevalence rate would have been if individuals had decided to have healthier lifestyles. In addition, I test a set of neighborhood amenities to see if they have any significant impact upon encouraging healthy behaviors, and the effect of this influence on obesity prevalence.

I find evidence of a significant reduction in the obesity prevalence rate for adult females and males derived from a hypothetical situation in which they perform intense physical activity when they are high school students. I also found evidence that a generalized, continuous practice of intense physical activity would produce big falls in the adult obesity prevalence rate. In addition, using the model estimation, I test if a set of neighborhood amenities has any significant impact upon the encouragement of physical activity. After controlling for the endogeneity of neighborhood amenities, I find that most neighborhood amenities are not significant in terms of encouraging residents' physical activity. Nevertheless, an increase in one standard deviation in the availability of a set of physical-activity-related amenities would produce a significant reduction of one percentage point in the adult obesity prevalence rate.

2 Background Literature

There is an increasing amount of literature about obesity in several disciplines of the social sciences; the recent interest in this topic has two main explanations. First, it is one of the most important public policy concerns in the US nowadays; according with the Center for Disease Control, it is the second leading cause for premature death, after smoking (Mokdad, Marks, Stroup, and Gerberding, 2000). The amount of resources spent every year in medical care of obesity and its consequences is high and has been increasing with the obesity prevalence rate (Wolf and Colditz, 1998; USDHHS, 2001; Bhattacharya and Sood, 2006; Folmann et al, 2006). The second reason is that the growth of obesity in the US during the last three decades was surprisingly high. Between 1960 and 1980 obesity prevalence rates in US were relatively stable (Rashad and Grossman, 2004), but after 1980 the prevalence rates more than doubled their original levels. Explaining this accelerated growth of obesity prevalence is a real puzzle and a very interesting question for many social scientists. In order to solve this puzzle, standard tools from economics could be very useful. This issue has opened a recent research interest that could be grouped under the terminology of "economics of obesity".

From an economic perspective, many elements could have contributed to the sharp growth in obesity. Recent literature has focused on the role that changes in relative price and cost of food may play. One of the first papers to formally state a hypothesis and prove some of its implications was Cutler et al. (2003), in which aggregate data was used to conclude that the recent obesity epidemic is explained by a reduction in the time cost of household meal production (as a result of this reduction, there has been an increase in the quantity and variety of food consumed). The technological change that characterizes post-industrial societies is another factor associated in the literature with the high obesity prevalence observed in recent decades (Lakdawalla & Philipson, 2002; Philipson & Posner, 1999). An example of this branch of the literature is Lakdalla, Philipson, and Bhattacharya's (2005) hypothesis that technological change has simultaneously lowered the cost per calorie and raised the cost of physical activity by making agricultural production more efficient and jobs more sedentary (Lakdalla et al., 2005). The contribution of technological change to obesity was also portrayed in Lakdawalla and Philipson (2002), a paper that presented evidence of the existence of an inverse relationship between Body Mass Index (BMI) and job strenuousness.

A recent series of papers has extended this research line by exploring the determinants of obesity and BMI as health outcomes derived from individual characteristics and choices. Usually these choices are variables that describe aspects of an individual's lifestyle (French et al., 2010, Grossman & Saffer 2004; Rashad 2006; Wen et al., 2010). In general terms, all of these papers specified obesity or BMI equations using micro-data. Therefore, in this sub-branch of the literature one could make the distinction between papers that control or do not control for potential endogeneity bias. Grossman and Saffer (2004), for example, did not address any possible endogeneity issue in regard to the elements they used as

explanatory variables. Rather than specific individuals' decisions, the variables they included in their BMI equations were prices and contextual variables that could be exogenous to some extent.

Other papers, such as Rashad (2006), French et al.(2010) and Wen et al. (2010) have specified the weight status equation in terms of individual decisions; in doing so, they assumed those decisions are endogenous and therefore implemented an empirical strategy to correct for the endogeneity bias. In the case of Rashad 2006, the author controlled for the endogeneity of smoking, calorie intake, and physical activity by using an instrumental variables estimation routine. French et al. (2010) took advantage of the longitudinal nature of their data and used fixed effect panel methods to get rid of time-invariant unobserved heterogeneity that could be correlated with the variables of interest. Wen et al. (2010) estimated a dynamic model of weight using longitudinal data from China. To control for the endogeneity of the individual choices included in the model (i.e., smoking, drinking, physical activity, and diet), they estimated a dynamic GMM system, using as instruments spatially varying macro-level factors such as urbanicity and prices.

Another branch of the literature has sought to identify the links between high levels of obesity prevalence and characteristics of the environments in which individuals live. The main hypothesis of the papers in this branch is that built environments may encourage individuals' physical activity and thereby have an impact on obesity. One variable that has caught the attention of many researchers is "urban sprawl," usually defined as the expansion of cities and their suburbs to rural areas. Researchers from numerous disciplines have tested the hypothesis of a significant relationship among urban sprawl, physical activity, and obesity (Ewing, Schmid, Killingsworth, Zlot, & Raudenbush, 2003; Giles-Corti, Macintyre, Clarkson, Pikora, & Donovan, 2003; Glaeser & Kahn, 2004; Lathey, Guhathakurta, & Aggarwal, 2009; Saelens, Sallis, Black, & Chen 2003). Other researchers have used wider measures of built environment beyond urban sprawl.

The built environment can be understood as a major component of community design; as such it is comprised of aspects such as buildings, transportation systems, parks, and greenways (Boone-Heinonen et al., 2009). Several papers have sought to measure the relationship between weight or energy expenditure measures and neighborhood density of physical-activity-related facilities (Boone-Heinonen & Gordon-Larsen, 2009; Gordon-Larsen, Nelson, Page, & Popkin, 2006). These works have stated a clear hypothesis, namely that neighborhood amenities and characteristics can improve population health by encouraging positive health habits in the community.

Some of the papers mentioned in the previous paragraph noted the importance of controlling for the endogeneity of neighborhood characteristics. Given the ability of households to choose a neighborhood that matches their interest in health issues, this variable should be treated as endogenous. In other words, the tendency of healthy people to look for healthy neighborhoods to live in can be interpreted

as a self-selection process. The usual strategy of some researchers in controlling for the endogeneity of neighborhood characteristics is to perform some kind of fixed effects estimator (Boone-Heinonen et al., 2010; Eid et al., 2008). A good example of papers based on fixed effects methodologies is Eid et al. (2008); they controlled for the endogeneity of neighborhood characteristics by using a first-difference estimator. Their results rejected the hypothesis of a significant relationship between urban sprawl and obesity. For a very comprehensive review of the evolution of this literature, the reader may refer to Boone-Heinonen et al. (2009).

The present research shares the common conception of several of the papers mentioned above, the idea that weight status can be modeled as a health production function that is determined by individual characteristics and choices. Some papers in the literature on obesity have focused on identifying the effect of lifestyle choices on weight status, whereas others have focused on identifying the effects of neighborhood characteristics (which are determined by the residential location decision) on weight status. In the present research I propose a comprehensive, dynamic model in which weight status appears as the combined result of lifestyle choices; at the same time, it recognizes that neighborhood amenities help determine the levels of physical activity that individuals decide to perform. Therefore, this dissertation can be seen as a bridge between these two types of obesity research. I control for the endogeneity of these choices by estimating the model jointly and by allowing errors to be correlated across equations. One of the choices that the individual is allowed to make in this model is the residential location decision. By explicitly modeling residential decisions I can control for the endogeneity of the neighborhood characteristics that are directly derived from it. The methodology itself is another contribution of the present research to the literature on obesity because in almost all cases, endogeneity issues have been controlled using Instrumental variables of fixed effects models. These approaches have some limitations, such as the impossibility of accounting for time-varying unobserved heterogeneity.

3 Data

The main source of information used in this study is The National Study of Adolescent Health (AddHealth). One of the main characteristics of this study is its comprehensive contextual information on the characteristics of the neighborhoods in which the respondents live. Because neighborhood characteristics are important in the present research, a subsection below is devoted to explaining the contextual information available in AddHealth and the definition of neighborhood used herein. A general explanation on the AddHealth study dataset is also provided.

3.1 The AddHealth Study

The National Study of Adolescent Health (AddHealth) is a longitudinal survey that began with a nationally representative sample of high school students in grades 7 to 12 during the 1994 and 1995 school

years. Respondents were followed after the first information collection and were interviewed three additional times. AddHealth explores adolescents' health-related behaviors and keeps track of them into young adulthood. Something unique about AddHealth, and very crucial for the present study, is that it contains very good information on the activities that respondents perform during their "active" leisure time. This information is important because the relationship between weight status and physical activity is where policy variables of interest may demonstrate influence. Other special features of AddHealth include its great diversity in terms of ethnic backgrounds (4,400 African Americans, 3,400 Hispanics, 1,500 Asians). In addition, it includes special oversamples of important populations such as sibling pairs and African-Americans with college-educated parents, as well as Cubans, Puerto Ricans, and Chinese. These special oversamples were also very important for the present study because they made it easier to identify and measure the existence of racial and ethnic health disparities..

The first wave (Wave I) of AddHealth, collected in 1995, consisted of 20,745 high school students in grades 7 through 12. The second wave (Wave II), collected in 1996, included 14,738 of the original Wave I respondents. The third wave (Wave III), collected in 2001, included 15,197 original Wave I respondents, most between 18 and 26 years of age. The last wave (Wave IV), collected in 2007, included 17,000 original Wave I participants between 24 and 32 years of age.

3.2 Contextual Information and Neighborhood Characteristics

A very important feature of AddHealth is the outstanding amount of contextual information it contains. This information describes a comprehensive set of characteristics of the environments in which AddHealth respondents live and is available for small areas, something that is not very common in public versions of longitudinal studies. Many variables are available at the Census tract level, and some are generated for even smaller geographical areas. An important subset of contextual variables, some of which are used in this dissertation, was generated to describe characteristics of an area equivalent to a buffer, with a specific radius whose center is the respondent's household (usually such buffers are defined with radii of 1, 3, 5, and 8 km).

Because many of the contextual variables used in the present research as explanatory variables have been generated at the census tract level, these areas are treated herein as one of the definitions for the individual's neighborhood. The following definition of census tract comes from the United States Census Bureau: "Census tracts are small, statistical subdivisions of a county. Census tracts usually have between 2,500 and 8,000 persons and, when first delineated, are designed to be homogeneous with respect to population characteristics, economic status, and living conditions. Census tract boundaries are delineated with the intention of being maintained over a long time so that statistical comparisons

can be made from census to census[2]". Some other contextual variables used in the present study were generated using the previously described buffer principle (the variables used as explanatory variables were generated for a radius of 5 km); therefore, the other neighborhood definition used here is a 5-km. buffer area with its center at the respondent's residential location.

In this research I am able to know the location of each responder through all the study. Therefore, I can know a set of characteristics of the neighborhoods where responders are located. Although contextual information is available for all waves of the AddHealth study, some variables are not available for all responders in the estimation samples in all waves. In order to be able to use the contextual information despite this missing values problem, I have performed imputations for some contextual variables. Details about these imputations, sources of the contextual data, and a general description of the variables in the estimation sample are provided in the data appendix.

4 Theoretical Motivation

The purpose of this section is twofold. The first subsection provides a static model in which the individual is allowed to make a set of decisions (e.g., choice of residence, education, smoking, physical activity, and food consumption). This simple model is useful as a way to theoretically identify the pathways through which weight can be affected by the set of choices that will be included in the empirical model. The second subsection outlines a dynamic framework that not only allows for the motivation of a potential specification for the empirical equations but also gives a theoretical justification for the sources of identification. In addition, the framework proposed in the second subsection is useful for studying the dynamics of weight status as a health outcome in a setting similar to the one proposed by Grossman (1972) in his seminal paper.

4.1 Simple Static Model

To illustrate how some individual's choices can determine weight status, a simple one-period model of weight, residence, and education can be useful. Let's assume the existence of a perfectly competitive market for housing, which is a differentiated product that can be completely described by a vector of objectively measurable characteristics $z = (z_1, z_2, ..., z_N)$; with z_i representing the amount of the characteristic i in the housing. Under standard assumptions, Rosen (1974) showed in his seminal paper that there exists an equilibrium in which implicit prices for each characteristic are derived $(p_{z_1}, p_{z_2}, ..., p_{z_N})$. These implicit prices are such that the amount of the characteristic z_i demanded by households exactly matches the amount of this characteristic supplied by the housing producers.

[2] This is a fragment of the official definition of a Census Tract offered in the Census Bureau Website www.census.gov/geo/www/cen_tract

In this simple model I make use of this powerful principle to represent the dwelling by a vector $z = \left(\vec{\eta}, \vec{\xi}\right)$, where $\vec{\eta}$ is the sub-vector of characteristics for housing that are somehow related to physical activity (e.g, urban sprawl, parks or recreation centers in the neighborhood, pool in the yard), and $\vec{\xi}$ is a sub-vector that collects all other characteristics of the dwelling. Assuming a hedonic equilibrium à la Rosen, the price of dwelling can be decomposed by a price vector that collects the implicit prices of each characteristic in each subgroup $p_z = \left(p_{\vec{\eta}}, p_{\vec{\xi}}\right)$.

The individuals are assumed to obtain utility from food consumption (f), smoking (s), their children (n), their dwelling in terms of its amenities (η,ξ) [3], leisure (l), a generic consumption good x, and their weight W. In addition, individuals in this model get utility from the intensity, in terms of energy expenditure, with which they spend their leisure time (ϕ). The term ϕ can be thought of as the whole amount of energy spent during their leisure time. In other words, individuals may choose how physically demanding their leisure activities are going to be and also get utility from the intensity with which they spend this time. The generic consumption good x is assumed to have no impact at all on the individual's weight.

Individuals have a fixed amount of time that they distribute to leisure (l), working (h), and acquiring education (e). Individuals cannot consume more than their income $[y(e) + y_o]$, which is represented as a function of the individual's education (e) plus initial wealth y_o. The possibility of credit markets is ignored. The utility function of the individual can be represented as

$$U[W, f, s, n, l, \phi, \eta, \xi, x] \tag{1}$$

In this model weight is assumed to be a function of food consumption (f), smoking (s), number of children (n) (for women only), and physical activity (a). All of these factors are under individual control and they have an impact on weight through biological process. The existence of relationships between weight and food consumption, and between weight and physical activity, is obvious. Smoking has been widely proven to have a negative impact on an individual's weight (Grunberg & Klein 1998, Flegal et al., 1995; Gerace et al. 1991; Green & Harari 1995; Mizoue et al., 1998; O'Hara et al., 1998). For women who have given birth, obesity may be due to weight retention after delivery. Several papers in the medical and epidemiological literature support this hypothesis, at least for some specific populations (Gunderson Abrams, 2000; Keppel & Taffel 1993; Ohlin & Rossner, 1990; Parker & Abrams, 1993; Rossner & Ohlin, 1995).

Physical activity, which plays a very important role in this model, is assumed to be a function of the time the individual spends in leisure (l) and working (h). The time spent in each activity is multiplied

[3] For simplicity in the notation η,ξ are scalar indexes that completely describe the whole variation in the vectors $\vec{\eta}$, and $\vec{\xi}$ respectively.

by efficiency parameters ϕ and θ respectively, which represent how demanding the activity in terms of physical effort is. Therefore, physical activity can be thought of as a measure of energy spent during the whole time with which the individual is endowed[4].

$$W = W(f, a, s, n) \tag{2}$$

$$a = \theta(\eta, e)\, h + \phi l \tag{3}$$

As previously explained, the parameter (ϕ) is an individual choice of leisure time energy expenditure. As such, it has a price as any other consumption good does; it can be thought as a measure of energy expenditure per unit of time. This measurement offers a way to model that individuals can perform different activities during their leisure time, that they know perfectly what the price of each one of those activities is, and that they know what their requirements are in terms of physical activity. The efficiency parameter (θ) is assumed to be a function of neighborhood amenities (η) and education (e). The intuition for this specification is that different careers imply different levels of energy expenditure. Education determines the occupations at which an individual can work; each one of those occupations represents a different level of energy expenditure. Neighborhood amenities can affect the levels of energy expenditure in labor activities, however, especially through the use of different transportation alternatives. For example, depending upon the neighborhood, individuals can take public transportation or bicycle to work.

The optimization problem that an individual solves is the maximization of equation (1) subject to (2), (3) and the following [5] budget and time restrictions (5) and (6):

$$x + p_f.f + p_s.s + p_n.n + p_\xi.\xi + p_\eta.\eta + [p_\phi - c(\eta)\,\tau].\phi = y(e) + y_o \tag{4}$$

$$l + h + e = \bar{T} \tag{5}$$

where $[p_f,\ p_s,\ p_n,\ p_\xi,\ p_\eta,\ p_\phi]$ represent prices for each of the consumption goods. The final cost of leisure energy expenditure (ϕ) depends of its price per unit p_ϕ, (which could be thought as the average price per calorie burned in market energy expenditure activities) and a negative cost function $c(\eta)$. This function represents the amount of the cost per calorie burned that can be avoided by substituting market energy expenditure activities for neighborhood amenities. For example, instead of using the treadmill in a conventional gym, individuals can run in the community park. Individuals in this model might also

[4] It is assumed that the time individuals spend acquiring education has no significant effect in terms of their energy expenditure.

[5] The price of the generic consumption good is normalized to 1 for convenience in the notation, all other prices and income are relative to the price of x.

decide how much of this cost reduction they are willing to take advantage of; in other words, conditionally upon the amenities of their neighborhood, they could decide how much market leisure energy expenditure they want to substitute. The parameter $\tau \in [0,1]$ could also be an individual decision; if the individual is willing to take advantage of all the amenities that her neighborhood offers, then τ will be close to one and she will get a great reduction in the final cost per calorie burned. If the individual does not want to take any advantage of her neighborhood amenities, then τ will be close to zero, and there will not be any reduction in the cost per calorie burned. For simplicity in the subsequent analysis it is assumed that $\tau = 1$.

Based on this simple model, one can explore from a theoretical point of view the nature of the relationships between individuals' consumption decisions and their weight. From the specification of the equations above one can see that some choices (e.g., smoking, food consumption, and family size) would directly affect the biological processes that determine an individual's weight status. Some other choices (e.g., education and leisure energy expenditure) would indirectly affect such biological processes by modifying an individual's levels of physical activity. Finally, the choice of neighborhood amenities would modify the final cost of each level of leisure energy expenditure, and in this way would affect the weight status by altering the rational levels of leisure energy expenditure chosen at these new prices. From the individual's optimization problem I am able obtain conditions that are informative about the endogenous nature of individual consumption choices. From the optimization conditions presented below, it is clear that when the individual rationally decides her consumption, there is a set of considerations that she will have to take into account. These considerations are directly or indirectly related to the determination of the weight status. Some of these optimization conditions can be seen in the following equations.

$$
\begin{aligned}
U_x &= \left(\frac{1}{p_f}\right)[U_f + U_W W_f] && (6{:}[f]) \\
&= \left(\frac{1}{p_s}\right)[U_s + U_W W_s] && (7{:}[s]) \\
&= \left(\frac{1}{p_n}\right)[U_n + U_W W_n] && (8{:}\,[n]) \\
&= \left(-\frac{1}{y'(e)\,h}\right)[-U_l + U_W W_a(-\phi + \theta_e.h)] && (9{:}\,[e]) \\
&= \left(-\frac{1}{y(e)}\right)[-U_l + U_W W_a(\theta - \phi)] && (10{:}\,[h]) \\
&= \left(\frac{1}{p_\eta - \phi c'(\eta)\,\tau}\right)[U_\eta + U_W W_a \theta_\eta h] && (11{:}\,[\eta]) \\
&= \left(\frac{1}{p_\phi - c(\eta)\,\tau}\right)[U_\phi + U_W W_a l] && (12{:}\,[\phi]) \\
&= \left(\frac{1}{p_s}\right)[U_\xi] && (13{:}\,[\xi])
\end{aligned}
$$

In the previous equations, Λ_r denotes $\frac{\partial \Lambda}{\partial r}$, with $\Lambda = \{U(.), d(.), W(.), a(.), \theta(.)\}$. This set of equations is based on the simple principle that individuals optimize their consumption when the marginal rate of substitution between two goods is equal to their relative prices. In this case I use the marginal rate of substitution between the generic consumption good (x) and any other good. These equations describe the fact that the individual's optimal behavior requires that the marginal utility derived from the consumption of one good, multiplied by the relative price ratio with respect to p_x, is equal to the marginal utility derived from the consumption of any other good (multiplied by the price ratio, with p_x normalized to 1).

Particularly, equations 6–8 define optimality conditions for the consumption of food (6), cigarettes (7), and children (8). From these equations one can see that the marginal utility from the consumption of these goods is composed of a pure physic-utility term and another term that always involves the partial derivative of utility with respect to W; I refer to this term as a "weight effect." The psychic-utility is the utility that people get directly from the consumption of a good; as such it excludes any other possible indirect effect through some other component of the utility function. The weight effect is a concept used in this dissertation to describe the indirect marginal effect that the consumption of some good has on the utility that an individual gets from his weight. From these first three conditions one can see that when individuals decide to engage in consumption of some goods, they will consider not only the direct utility they get from this consumption but also the ultimate implications that this consumption has on their weight, weighted by the marginal utility of an additional pound. For example, people smoke because they like cigarettes, but also because they may like the effect that smoking has on their weight. This dual attraction implies the existence of a reverse causality that will be a source of endogeneity. Smoking has an impact on weight, but at the same time, unobservables driving the preferences about W make individuals prone to smoking.

In the remaining conditions (9–13) a similar interpretation applies: weight effect may play an important role when individuals are making optimal choices about education, labor supply, neighborhood amenities (physical-activity-related), and leisure energy expenditure. In the case of education, for example (9), the second term inside the brackets describes two effects related to weight. The first is the effect of the foregone leisure that could have implied increments in physical activity and thereby in weight changes. The second is the effect of education on the efficiency parameter θ. This second effect is mainly driven by the term θ_e, which describes the change in energy expenditures per amount of time dedicated to work when education is marginally increased. When individuals in this simple model make choices about education, they have already taken into account that energy expenditure levels vary across the activities and occupations they will perform, given the education they decide to acquire.

In the case of amenities (η) (11), this choice will have an indirect weight effect through the changes in the efficiency parameter θ, which determines energy expenditure in labor activities (through local

transportation facilities, for example). In addition, the choice of (η) will modify the relative prices in equation (12); in other words, it will modify the ultimate cost of leisure energy expenditure, which in turn implies changes in physical activity and weight. To summarize, conditions 6–13 reveal a reverse causality between obesity and factors that intuitively may explain it. When people decide in favor of the consumption of some good that can be associated as a causal factor of obesity, the optimal consumption of that good is partly explained by the preferences that individuals have about their weight.

4.2 Dynamic Framework

The following dynamic framework reflects the intuition behind the one-period model; in order to avoid non-essential over-parameterization, however, some simplifications are implemented. This framework is more adequate for introducing the empirical model because individuals in AddHealth are observed several times during a relatively long period of their lives. In this section of this dissertation, the idea of education as a choice variable is extended to more general career-related decisions. The AddHealth respondents are undergoing major life transitions and a significant number are in college, leaving college, working, in vocational schools, finishing high school, in the military, and even in prison. In order to take advantage of that information, career decisions instead of educational choices are considered here. Within this dynamic framework, individuals are allowed to make choices about their careers and residential locations; similarly, they make decisions about food consumption, fertility, and smoking; and they decide the intensity of physical activity during leisure time. Finally, as a result of all these choices and the relationships among them, the weight of the individual is produced as a health outcome. The timing assumed for this decision process is explained below.

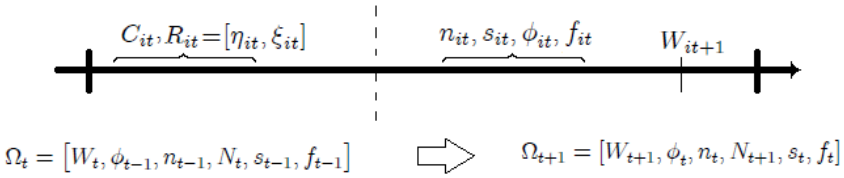
An important feature of this model is its dynamic nature, the model is dynamic in the sense that previous behavior influence current decisions. This distinction of the model is important in the theoretical framework and in the empirical model. The theoretical justification for this comes from the traditional theory of rational addiction (Becker and Murphy, 1988). The theory of rational addiction suggest that utility of an addictive good is influenced by previous consumption behavior. The rational addiction framework have been traditionally used to model risky behaviors as smoking (Gilleskie and Strumpf, 2005; Chaloupka, 1991; Labeaga, 1999), but it can be extended to more general demand decision in which state dependence may play a role. In this research state dependence is understood as the situation in which previous consumption of a specific good has a significant impact in its current consumption (Gilleskie and Strumpf, 2005).

Timing Assumptions

Timing assumptions will play an important role in the specification of the empirical model. These assumptions are summarized in figure 1. The information with which an individual enters at period t

is stacked in the vector $\Omega_t = \left[W_t, \phi_{t-1}, n_{t-1}, N_t, s_{t-1}, f_{t-1}\right]$. This information includes the weight at the end of the previous period (or beginning of the current one) W_t, the intensity of physical activity in the previous period ϕ_{t-1}, the fertility decision from the previous period n_{t-1} and the family size at the end of previous period N_t, as well as the food consumption f_{t-1} and the smoking indicator s_{t-1} from the previous period. After considering this information, individuals simultaneously make the first two decisions in the period. The career related decision c_{it} and the residential decision, which is represented by the characteristics of the dwelling - including neighborhood amenities - $R_{it} \equiv [\eta_{it}, \xi_{it}]$. Next, individuals make the following four simultaneous endogenous choices for the current period: food consumption f_{it}, number of children in this period n_{it}, smoking decision s_{it}, and intensity of their physical activity in the current period ϕ_{it}. The weight at the end of the period W_{t+1}, is determined by the weight at the beginning of the period W_t, and the endogenous choices made within the period. Based on behaviors during period t, the vector of state variables Ω_t, evolve to the next period $\Omega_{t+1} = [W_{t+1}, \phi_t, n_t, N_{t+1}, s_t, f_t]$.

Figure 1: Timing Assumptions



An implicit consideration in this time framework is important for identification purposes, namely that shock prices that affect the choices of $[n_{it}, s_{it}, \phi_{it}, f_{it}]$ in the second intra-period stage occur after the first intra-period stage choices are made. Therefore, individuals learn about these shocks after they have made the residential location and career-related decisions. These shock prices at the beginning of the second intra-period stage could be interpreted as new information that appears between intra-period stages. The intuition for this assumption is that career-related and residential choices are two major decisions in the individual's life. Depending upon their decisions about career and residence, individuals will end up in locatios with different distributions for prices. These shock prices are stacked in the vector $\Psi_t = \left[p_{\eta_t}, p_{s_t}, p_{\phi_t}, p_{f_t}\right]$.

Theoretical Framework

Individuals derive utility from leisure time energy expenditure (ϕ_{it}), smoking (s_{it}), food (f_{it}), newborns (n_{it}), total family size (N_{it}), dwelling characteristics (including neighborhood amenities) $[\eta_{it}, \xi_{it}]$, leisure (l_{it}), and the composite consumption good (x_{it}). Individuals also get utility from their weight (W_{it}), which they cannot decide but can control by making the set of choices described above. The utility function also depends upon individual exogenous characteristics X_{it}, and an unobserved component u_{it} that can be thought as a standard preferences shock. The utility function of the individual i in the

period t can by represented as

$$U_{it} = U\left[W_{it}, f_{it}, s_{it}, n_{it}, N_{it}, l_{it}, \phi_{it}, \xi_{it}, \eta_{it}, x_{it}, u_{it}; X_{it}\right] \tag{2.1}$$

To avoid unnecessary complications, in this model I collapse education decisions, and labor supply decisions. All that information is contained in the career-related choices that an individual makes throughout her life. Human capital in this framework $\left(H_{it}^{k}\right)$ is the experience and education accumulated in a specific career $k \in \{1, 2, ..., K\}$. The accumulation of human capital depends on current and previous career decisions $(c_{it}, c_{it-1}, ..., c_{io})$ and its evolution is explained later. Individuals may accumulate human capital in several careers; their income will depend on the amount of human capital accumulated in each one of them.

Career-related decisions also determine the individual's time allocation. Individuals in this framework decide their careers, and each career has its own time requirements in terms of labor supply (h_{it}) and school attendance time (e_{it}). In other words, the individual's time allocation is modeled as the result of the career decision rather than a choice by itself. This way of modeling individuals' decisions (time allocation and human capital investment) is convenient for the purposes of this study because different careers can be easily associated with different levels of energy expenditure. The time constraint will be represented as:

$$T = l_{it} + h_{it}\left(c_{it}^{k}\right) + e_{it}\left(c_{it}^{k}\right) \tag{2.2}$$

Before deciding their career-related choice (c_{it}) and residential location (η_{it}, ξ_{it}), individuals observe the information available at the beginning of the period, which is includes the values of previous choices and the previous realization of the health outcome. After these first two major decisions are made, individuals learn the characteristics of their residential location, and their career choice, and they observe the price shocks. Then they make the remaining decisions $(s_{it}, n_{it}, \phi_{it}, f_{it})$. At the end of the period, and as a result of the influence of all the endogenous choices, the health outcome - weight status - W_{it+1} is produced. In short, during each period the individual makes residential, career-related, leisure energy expenditure, smoking, fertility, and food consumption decisions. For convenience in the notation the choice variables in each period $\{s_{it}, n_{it}, \phi_{it}, f_{it}, \eta_{it}, \xi_{it}, c_{it}\}$ are grouped in two vectors. One represents lifestyle decisions $l_{it} = [s_{it}, n_{it}, \phi_{it}, f_{it}]$, and the other represents major individual decisions $m_{it} = [\eta_{it}, \xi_{it}, c_{it}]$.

The total number of children at the end of the period is determined by the fertility decision during the current period (n_{it}) plus the total number of children accumulated from the previous period; this process is described by Equation 2.3. As previously mentioned, human capital is modeled in this study as the accumulation of education and experience in a specific career; it is determined by the current

career decision c_{it}^{k}, and the amount of human capital accumulated until to the previous period in the same career k. This human capital accumulation process is described in Equation 2.4. The function $\varphi(.)$ maps the current career decision into human capital; for the purposes of this study, the specific unit of measure does not matter. Finally, the weight at the end of the current period is the result of a biological process of energy intake and energy expenditure, in which additional physiological factors play a role. This process can be represented by Equation 2.5, which represents the process of individuals' weight determination, which in turn depends upon an individual's weight in the previous period plus a series of inputs that represent individual choices. In other words, Equation 2.5 represents the final effects on the individual's weight of the following factors: energy intake, energy expenditure, and physiological alterations derived from individual choices. Number of childbirths during the period is considered in this model as a weight determinant exclusively for women; the intuition behind its inclusion is based on the hypothesis of weight retention mentioned in the previous section.

$$N_{it+1} = N_{it} + n_{it} \tag{2.3}$$

$$H_{it+1}^{k} = H_{it}^{k} + \varphi\left(c_{it}^{k}\right) \quad \forall\ k = 1, ...K \tag{2.4}$$

$$W_{it+1} = W(W_{it}, a_{it}, s_{it}, n_{it}, f_{it}) \tag{2.5}$$

The total amount of physical activity per period (a_{it}), as in the static model, is a measure of energy spent during the whole period. In this framework it is assumed to be a function of the environment in terms of neighborhood amenities, career-related choice, leisure energy expenditure levels, and the time allocation. These arguments are the same primitive factors that explain the energy expenditure in the static case. Therefore, the physical activity that an individual undertakes in period t can be represented as:

$$a_{it} = a\left[\eta_{it}, c_{it}, \phi_{it}, l_{it}, h_{it}\right] \tag{2.6}$$

The individuals in this theoretical framework solve a dynamic optimization problem subject to all previous equations and constraints, plus one additional budget constraint. The total income is a function of the labor supply implied by the career decision c_{it}^{k} multiplied by an income function $y(.)$, which depends on the total human capital accumulated in the same career k and a basic level of human capital H_o. This basic level of human capital is a threshold that can be obtained from the accumulation of human capital in any career. The budget constrain can be represented as

$$
\begin{aligned}
p_x x_{it} + p_f f_{it} + p_s s_{it} + p_n n_{it} + p_\xi \xi_{it} + p_\eta \eta_{it} + [p_\phi - c(\eta_{it})]\,\phi &= y\left(H_{it}^k, H_o\right) h_{it}(c_{it}^k) \\
\forall\ k &= 1, 2, ..., K \qquad (2.7) \\
\text{Where}\ \ H_o &= \begin{cases} 1 \text{ if } \sum_k H_{it}^k \geqslant \bar{H} \\ 0 \text{ if } \sum_k H_{it}^k < \bar{H} \end{cases}
\end{aligned}
$$

Additional characteristics of the budget constraint are the same as in the static model (for further explanation, please refer to the previous section). At any period t, the objective of the individual in this model is maximize the expected present discounted value of the remaining lifetime utility.

$$
E_t\left[\sum_{\tau=t}^{T} \beta^{(\tau-t)} U\left(W_{it}, f_{it}, s_{it}, n_{it}, N_{it}, l_{it}, \phi_{it}, \xi_{it}, \eta_{it}, x_{it}u_{it}; X_{it}\right)\right] \qquad (2.8)
$$

subject to (2.2 and 2.7)

Where β represents the discount factor. A sequential representation of this lifetime discounted utility optimization problem can be made using a Bellman equation for any combination of choices $[\phi_{it}, \tau_{it}, s_{it}, n_{it}, \eta_{it}, \xi_{it}, x_{it}]$ with the following value function:

$$
\begin{aligned}
V_{l,m}\left(X_t, \Omega_{it}, u_{it}, \Psi_t\right) &= U\left(W_{it}, f_{it}, s_{it}, n_{it}, N_{it}, l_{it}, \phi_{it}, \xi_{it}, \eta_{it}, x_{it}, u_{it}; X_{it}\right) \qquad (2.9) \\
&\quad + \beta . E\left[V\left(X_{it+1}, \Omega_{it+1}, u_{it+1}, \Psi_{t+1}\right)\right]
\end{aligned}
$$

where,

$$
V\left(X_{it+1}, \Omega_{it+1}, u_{it+1}, \Psi_{t+1}\right) = \max_{l,m}\left\{V_{l,m}\left(X_{it+1}, \Omega_{it+1}, u_{it+1}, \Psi_{t+1}\right)\right\}
$$

The expectation in Equation 2.9 is taken over the distribution of the random components that determine individual choices (in this case, the price shocks and the preference shocks). Demand functions for the choice variables result from the solution to this optimization problem. Substituting these demand functions into the health production function yields an expression for the weight status function. Approximations for all these equations are estimated in the empirical model.

(Chapter head:)Empirical Model

The coefficients of the health production function will be inconsistently estimated using standard methods such as OLS or an independent discrete outcome models. The weight outcome is a function of the individual's choice variables, and the endogeneity of these choices should be taken into account in the estimation. Variables such as smoking, number of children, and physical activity are endogenous because they are choice variables such that their optimal consumption depends somehow on the final

indirect effect that they have on the individual's weight status; furthermore, unobservables explaining each one of these behaviors may be conrrelated each other. In addition, neighborhood amenities are endogenous to the physical activity decision because individuals may choose their place of residence as a response to these amenities as well as their potential effect on their health (i.e., healthy people look for healthy neighborhoods). This is presented as a standard selection problem. Another issue that must be considered is that AddHealth respondents are undergoing major life transitions. These transitions complicate the estimation because career-related decisions should be included as an important element, but this choice is also endogenous as discussed in section four.

The timing scheme described in figure 1 assumes that when individuals make decisions, they will consider all the information available at the moment of the decision. The available information is composed of previous choices and stated variables at the beginning of the period. As mentioned before this is consistent with the standard framework of rational addiction. This sequence implies that the model is dynamic and that part of the identification will be based on this dynamic nature. In addition, the empirical model I include individual unobserved heterogeneity in each of the equations. This is important for two reasons. First, it allows modelling unobserved factors (e.g., preferences) that could be sources of endogeneity. Second, it provides a flexible way to model the correlation of unobserved factors across equations. In order to do so, however, some restrictions must be imposed on the distribution of the error terms; these restrictions are explained below.

4.3 Error Structure

Observed factors do not explain all variations in each of the outcomes and choices modeled in this dissertation; unobserved characteristics may also determine each one of these behaviors, and these unobserved characteristics may be correlated across equations. Consider, for example, unobserved preferences about physical activity; individuals who enjoy physical activity may participate in sports and outdoor activities, walk to work, or use public transportation. When this type of individual makes residential decisions, she is likely to choose neighborhoods with a set of amenities that would allow her to perform these kinds of activities. In order to take into account these correlations, I estimate weight status, career decisions, residential decisions, fertility decisions, smoking, food consumption, and physical activity jointly rather than separately. In addition, a flexible structure is imposed in the distribution of unobservables; this allows correlation among the different equations.

The correlations patters are modeled by decomposing the error terms of each equation into three parts $(\varepsilon_{it}, \mu_i, \upsilon_{it})$. First is an independent and identically distributed component which is assumed to be a type 1 Extreme Value or normal distributed error (ε_{it}) that can be interpreted as an idiosyncratic shock. The second and third components represent permanent (μ_i) and time varying (υ_{it}) unobserved

individual characteristics. I denote each one of the equations in the system by $e = \{1, 2, ..., 7\}$, and the total error by ϵ_{it}. This decomposition allows for nonlinear unobserved heterogeneity components in the total error structure. More specifically,

$$\epsilon_{it}^{e} = \mu_{i}^{e} + \upsilon_{it}^{e} + \varepsilon_{it}^{e}$$

One intuitive way of thinking about the unobserved heterogeneity parameters is the following: There are different types of individuals in terms of unobserved factors that researchers cannot observe; these include preferences and tastes, personality traits, and so forth. There is a distribution of these types of individuals in the population, and for each type of individual unobserved heterogeneity parameters differently affect their consumption decisions. Nevertheless, these unobserved heterogeneity parameters are correlated among different equations that explain individuals' behaviors. In order to estimate these unobserved heterogeneity parameters and the joint distribution of the parameters in different equations, I use a semi-parametric discrete factor approximation method. The Discrete Random Method is more general than other methodologies than assume an arbitrary distribution for the unobserved heterogeneity (Heckman & Singer, 1984). The cumulative distribution of the unobservable factors is approximated by a step function with a finite number of points of support, and the values and heights of the points of support are parameters estimated simultaneously with the other parameters of the model (Mroz, 1999; Angeles, Mroz, & Guilkey, 1998). The joint distribution of the unobserved effects is modeled as a multivariate discrete distribution with several points of support and is estimated jointly with all other parameters of the model. A more detailed discussion about unobserved heterogeneity parameters estimation is provided at the end of this chapter.

4.4 Empirical Equations

4.4.1 Residential Location Decisions

Mixed Logit Specification The first decision incorporated in the system is the place of residence at waves III and IV. In the theoretical in section (4.2) it was assumed that the individual makes her residential location decision and career-related decisions simultaneously. In a reduced-form environment, it is not possible to specify a multinomial logit for careers that includes the whole information set from which the residential location decisions is made. Approximations can be made by combining the two decisions, but to do so would heavily increase the number of parameters to estimate; nor would such a model be easy to interpret. Therefore, in this empirical approximation to the theoretical model I assume that individuals make these decisions sequentially (i.e., first they decide their residential location and then their career).

Location might play a role in the determination of weight status through the encouragement of

physical activity. At the same time, neighborhood amenities that are resultant from residence location decisions may be endogenous variables in an equation that explains physical activity. This is because the residence location might be partly explained by the good health practices of respondents (e.g., as physical activity). In this dissertation the equation for residential decisions is modeled as a mixed logit, which technically is a conditional logit augmented with non-alternative varying characteristics. The main feature of a conditional logit is that the regressors vary across alternatives. In other words, the latent utility level for each choice is assumed to be a function of the attributes of each alternative, as one would expect from a residential location decision.

A very special feature of the conditional logit model is that it can allow individual specific characteristics (not varying by alternative) that have a separate effect on the utility level of a specific choice. This type of specification is usually called a mixed logit. In fact, the standard multinomial logit can be expressed as a specific case of the conditional logit (Cameron & Trivedi, 2005; Wooldridge, 2007). The specification of the residential choice as a mixed logit is very convenient for the present research because it allows the latent utility level of the alternatives to vary with the characteristics of the neighborhood; in addition, it allows the inclusion of individual regressors (here, the same ones will be used in the estimation of the choice model for career-related decisions). Under standard assumptions about the distribution of the error terms, the probability that individual i at time t will choose the alternative k, in the mixed logit model, can be represented as:

$$\begin{aligned} P\left(R_{it}=k\right) &= P_{it}\left(k\right)=\frac{\exp\left(V_{itk}+\sum_{l\in K^R} d_{itkl}.V_{it}\right)}{\sum\limits_{k'\in K^R}\exp\left(V_{itk'}+\sum_{l\in K^R} d_{itk'l}.V_{it}\right)} \\ d_{itkl} &= \left\{\begin{array}{c} 1 \text{ if } l=k \\ 0 \text{ if } l\neq k \end{array}\right., \; k\in K^R, \text{ and } K^R=\{1,2,3,...,R\} \end{aligned} \tag{5.2.1}$$

where V_{tk} is a linear function of the alternative varying regressors, and V_{it} is a linear function of the individual specific, not alternative varying, regressors. The parameter d_{itkl} represent a dummy variable that is equal to one when $l=k^r$. Note that if $d_{itkl}=0$ for all l this model is a standard conditional logit. The original choice set K^R is the total set of neighborhoods from which the individuals in this model can choose. For the AddHealth respondents, this would be a very large number of alternatives.

After replacing V_{itk} and V_{it} with their parametric representations, the log of the probability ratio between location k^r and location 1 for the conditional logit can be written as:

$$\ln\left[\frac{P\left(R_{it}=k^r\right)}{P\left(R_{it}=1\right)}\right]=\left(Z_{itk^r}-Z_{it1}\right)\beta^R+X_{it}\left(\gamma^R_{k^r}-\gamma^R_1\right)+\;\Omega_{it}(\phi^R_{k^r}-\phi^R_1)+\mu^R_{i,k^r}+v^R_{it,k^r} \tag{5.2.2}$$

where

$$\Omega_{it} = [W_{it},\ A_{it-1},\ S_{it-1},\ n_{it-1},\ f_{it-1}];\ t = 3, 4;\ k^R \in \{2, .., R\}$$

The choice set of all possible alternatives turns out to be prohibitively large (over 2000 different alternatives). Because this large number of alternatives in the choice set would make the estimation intractable, it is necessary to reduce the number of alternatives from which the individuals are allowed to choose. In this study I use two different techniques to deal with this problem. The first is based on random sampling of the choice set for each individual, and the second is based on the aggregation of alternatives by types of neighborhoods. More detailed explanations of each method are provided in the next subsection.

The vector Z_{itj} in Equation 5.22 collects a set of location-specific variables or amenities in location j. In the specification, I also include individual-specific regressors similar to the ones included in the vector of exogenous individual characteristics X_{it}, and previous realizations of endogenous characteristics Ω_{it}. The vector Ω_{it} contains the individual's weight status (W), physical activity (A), smoking decision (S), fertility decision (n), and food consumption (f), all from the previous period. The error structure allows for time-invariant and time-varying unobserved heterogeneity terms. Readers should note that I have specified a set of permanent and time-varying unobserved perturbations per category; in other words, this specification allows for unobserved heterogeneity controls (i.e., unobserved preferences shocks) per each neighborhood in the individual's choice set.

Dealing with the Extremely Large Choice Set Problem The first method I use to deal with the intractable choice set was an aggregation of the categories into different "neighborhood types." Each type is a new aggregated category with characteristics equal to the mean characteristics in the specific type. The final choice set is formed by the chosen neighborhood, which also represents the type it belongs to, and the remaining aggregated categories. In order to define different types of neighborhoods, I use a non-hierarchical cluster analysis method[6] to form clusters based on different neighborhood characteristics. Each cluster defines a different type of neighborhood based on the characteristics and amenities of the neighborhood. In order to generate the cluster for type of neighborhood, I use a partition cluster methodology for a pre-established partition of five groups. The variables included for the generation of clusters were: the proportion of neighborhood population with a bachelor degree (or more), the median neighborhood family income, the neighborhood population density, the arrest rate per 100,000 neighborhood inhabitants, the number of colleges less than 5 km from any neighborhood border, the number of shopping centers less than 5 km from any neighborhood border, the number of points of

[6] A standard problem in social sciences, is collapsing the information contained in several varables into a single one. One of the methodologies most used by researches facing this problem is called cluster analysis, a term that encompasses a broad set of methodologies that are designed to extract the structure that fits better with the nature of the data. Kaufman and Rousseeuw (1990) defined cluster analysis as "the art of finding groups in data." These techniques are useful for organizing data by grouping objects of a similar kind within a finite amount of different categories.

interest (museums, theaters, etc.) less than 5 km from any neighborhood border, and the number of some additional physical-activity-related facilities in the neighborhood. Additional details on the cluster procedure and a table with summary statistics of neighborhood characteristics by cluster are provided in Appendix F.

The second method I use to deal with the intractable choice set was a random sampling of the choice set per individual. Under some minimal conditions this technique has been proven to provide consistent estimators that use a subset of the choice set (McFadden, 1978)[7]. The idea of the methodology is to use a subsample instead of the whole set of alternatives available for each individual; it has been used in several papers about residential location decisions and other individual decisions made from a huge choice set of alternatives (Liu et al., 2010; Parsons & Kealy, 1992; Train et al., 1987). Following McFadden's (1978) original notation, in this dissertation the subsample of neighborhoods available for each household i is denoted by D. The conditional probability of assigninig a subsample D to an individual i, given that k is the neighborhood chosen at time t will be denoted as $\Pi\left(D|R_{it}=k\right)$; with R_{it} denoting the individual's decision.

The conditional probability that individual i chooses neighborhood k at time t conditional on the sample of alternatives D (applying Bayes theorem) would be:

$$P\left(R_{it}=k^{r}|D\right)=\frac{\Pi\left(D|R_{it}=k\right).P(R_{it}=k)}{\sum_{j\in D}\Pi\left(D|R_{it}=j\right).P(R_{it}=j)} \tag{5.2.4}$$

By substituting Equation 5.2.1 into Equation 5.2.4, I got the following expression:

$$P\left(R_{it}=k^{r}|D\right)=\frac{\exp\left(V_{tk}+\sum_{l\in K^{R}}d_{itkl}.V_{it}+\ln\Pi\left(D|R_{it}=k\right)\right)}{\sum_{j\in D}\exp\left(V_{tj}+\sum_{l\in K^{R}}d_{itjl}.V_{it}+\ln\Pi\left(D|R_{it}=j\right)\right)} \tag{5.2.5}$$

The reader may note that this expression is almost identical to the expression for the probabilities in a standard mixed logit model, as in Equation 5.2.1, but it includes a correction parameter $\ln\Pi\left(D|R_{it}=k\right)$. The correction parameter is the log of the conditional probability of drawing subsample D given the choice k. In other words, $\Pi\left(D|R_{it}=k\right)$ is the conditional density driving the sampling procedure; conditionally on k, it tells us the probability that a subsample D is assigned to individual i. MacFadden (1978) showed that, under a minimal condition called "positive conditioning property," the maximum likelihood estimators of a model with probabilities given by Expression 5.2.5 would be consistent estimators. Furthermore, under a more restrictive condition, called "uniform conditioning property" the likelihood function obtained from expressions (5.2.5) and (5.2.1) will be the same.

[7] The consistency of the estimation using random sampling has been proved in the context of standard multinomial logits (McFadden, 1978).

In this study I implement a random sampling procedure that is consistent with the estimation of the mixed logit. The mixed logit can be seen as having been formed by two components: one conditional component (e.g., regressors that vary across alternatives) and one un-conditional component (e.g., regressors that are invariant across alternatives). In order to allow the identification of the un-conditional part, a normalization of parameters is needed as a consequence of the restriction that probabilities sum to one. Because this normalization requires the definition of a reference category, the random choice subset must conserve a structure that allows the definition of this reference category. The definition of type of neighborhood based on the previously explained cluster analysis provides a good structure of partitions of the whole choice set, which will be useful for the implementation of the sampling procedure.

The random sampling procedure implemented for the construction of the choice subset is based in one of the examples described in McFadden (1978)[8]. The choice set K^R is partitioned into sets $\{C_1, C_2, ..., C_L\}$ with E_l representing the cardinality of set C_l. The choice subset will be formed by the gathering the chosen neighborhood k from partition set C_k with one randomly selected alternative from each remaining partition set. This type of procedure holds the "positive conditioning property"[9]. The previously described sampling procedure requires the partition of the choice set K^R. A common practice in the literature (Chattopadhyay, 2000; Friedman, 1975), especially in models of residential location demand, is to limit the number of alternatives in the analysis by partitioning the choice set by community and major dwelling type (McFadden, 1978). After this partition is implemented, the random sampling is performed within each of the partitions. In this study I use partitions defined as types of neighborhoods. Instead of arbitrarily defining different types of neighborhoods, I use the non-hierarchical cluster procedure described at the beginning of this subsection to form clusters based on different neighborhood categories. Each cluster defines a different type of neighborhood based on the characteristics and amenities that the neighborhood has.

4.4.2 Career Related Decisions

What is a Career Choice in this Study? The career-related decision is an endogenous factor that plays a role in the determination of subsequent decisions. In addition, as suggested by the theoretical motivation, the career will also have a direct impact on weight status. Career-related decisions are taken as exogenous until Wave III; this is because practically all of the respondents in the sample are high school students up to wave II. One of the empirical challenges of this study has been defining a "career-related" decision. Not only can many factors contribute to the classification of an individual in a specific career, there can also be several dimensions of the same concept. For the purposes of this dissertation,

[8] Example (c-4) subsection 7, MacFadden(1974)

[9] As noted in McFadden (1978) the conditional probability of drawing a subsample D given the choice k can be represented as $\pi(D|R_{it} = k) = \frac{E_k}{\prod_{l=1}^{L} E_l}$ if $k \in D$, and $D \cap C_l \neq \phi$ for $l = 1, 2, ..., L$

a definition based merely on educational aspects (e.g., a bachelor's major) will not be enough. In order to analyze the relationship between career choice and health outcomes, a broader understanding is required. Therefore, in this study a career-related decision involves several aspects of the individual's life: educational decisions, vocational education, labor supply and occupational characteristics, involvement with the crminal justice system, and military service.

Defining Career Decisions Using Cluster Analysis In order to create different, mutually exclusive categories that represent a career decision, I cluster individuals using the information contained in the five categories mentioned in the previous subsection. Technically, I perform a partition clustering methodology for observations[10]; this method allows the individuals to be classified into several different groups without overlap. The cluster methodologies used for non-hierarchical partitions usually require an ex-ante number of clusters that have been predetermined by the researcher. The cluster analysis identifies groups of individuals that share very distinctive characteristics that are also visibly different from characteristics of other clusters. Using the most remarkable differences between groups, I design categories that capture these differences among clusters.

Appendix E contains a table that presents summary statistics of characteristics for each cluster obtained from a k-median cluster procedure with five predetermined categories. This k-median cluster procedure gathered individuals with similar characteristics that describe choices on the life-paths that individuals have decided to follow; therefore, construction of the categories for career-related decisions was based on the characteristics of the clusters. The career-related decisions are defined here in terms of six categories. An individual is observed at any of these categories at a specific period of this study:

In *college (or similar).* Most of individuals in this group reported having some college or more educational attainment. They are not full-time workers; in fact, most of them do not work. This group's most distinctive characteristic is that all of the individuals reported that they are attending school regularly. Most of them attend college or universities; therefore, the educational attainment for most of them is at least some college. A small share of individuals have not yet received a high school diploma, however, or reported receiving vocational education.

In a high skilled white-collar job. These are full-time/part-time workers with white-collar jobs who are not attending school; their educational attainment is a college degree or more[11].

[10] Usually such methods are grouped under the name "k-mean" or "k-median procedures." The basic idea was to create a predetermined number of clusters by an iterative process in which an observation k is assigned to a group with a close mean or median to the characteristics of the observation. Based on this iterative process, the new means/medians for each group are created. The process is continued until no observation is assigned to a different group. For a better understanding of this methodology, see Kaufman and Rousseeuw (1990).

[11] A small subgroup of individual highly educated (with college degree or more), which reported not having a job at the moment of the interview were added to this second category

In a high skilled blue-collar job. These are full-time/part-time workers with blue-collar jobs who are not attending school; their educational attainment is an associates degree, some college, or more.

In a low skilled white-collar job. These are full-time/part-time workers with white-collar jobs who are not attending school; their educational attainment is some college or less.

In a low skilled blue-collar job. These are full-time/part-time workers with blue-collar jobs who are not attending school; their educational attainment is some college or more.

Not working nor in school. These individuals reported school attainment of less than high school, high school, vocational degree, or associates degree; in addition, they reported that at the time of the interview they were not working or attending school.

These six categories describe a period t career-related decision that these individuals are following. I generate them based on the following individual characteristics: educational attainment, labor supply, characteristics of the main job, formal education institution attendance, and vocational education. Some other characteristics were used in the cluster analysis that these divisions were based on, but the features listed here turned out to be the most important determinants of the inclusion of an individual in a specific cluster. Using these categories, I was able to specify a discrete choice model of career-related decisions.

Career Choice Equation Denoting C_{it} as the career related decision (CRD) of individual i at period t (defined as defined in the previous sub-section), the following equation describes the log of the probability ratio between choice k^c and choice 1 where $k^c = \{$CRD category 1, CRD category 2, ..., CRD category $K^c\}$.

$$\ln\left[\frac{P\left(C_{it}=k^c\right)}{P\left(C_{it}=1\right)}\right] = Z_{it}\beta^C_{k^c} + X_{it}\gamma^C_{k^c} + \Omega_{it}\phi^C_{k^c} + \mu_i^{k^c} + \upsilon_{it}^{k^c} \tag{5.1}$$

$$\text{Where } \Omega_{it} = [W_{it},\ A_{it-1},\ S_{it-1},\ n_{it-1}, f_{it-1}];\ t = 3, 4;\ k^c = 1, ..., K^c.$$

The matrix X_{it} includes a set of individual exogenous variables affecting the choice; this matrix contains individual specific variables such as age, family background, and socio-economic characteristics. Z_{it} is a vector of characteristics of the individual's place of residence. In the theoretical framework it was assumed that career-related decisions are made after an individual observes the information available at the beginning of a period (Ω_{it}); this vector Ω_{it} contains the person's weight status, physical activity, smoking decision, and fertility decision, all of them at the previous period. The error term is composed of $\mu_i^{k^c}$ and $\upsilon_{it}^{k^c}$; where $\mu_i^{k^c}$ represents unobserved individual level characteristics that are constant over time, and $\upsilon_{it}^{k^c}$ represents time varying unobserved individual characteristics. These unobserved heterogeneity

parameters are allowed to change by category. In addition, a purely random type I extreme value error perturbation is implicit in the multinomial logit specification.

4.4.3 Simultaneous Choices and Final Health Outcome

It was assumed in the previous section that after the residence and career choices, individuals make four simultaneous decisions: physical activity, smoking, number of pregnancies, and food consumption. Final health status is realized at the end of the period, after these simultaneous decisions are made. This assumed decision process describes the intuition behind the fact that individuals cannot decide their weight, but they can make their choices in such a way they can control their final weight status through these choices.

There is no information in Addhealth that allow me to construct a measure of caloric intake. This is a limitation of most of the papers on the relationship between obesity and environment. The implications of the lack of caloric intake information may be attenuated by the fact that there is neither evidence of variation in diet across physical activity levels nor evidence that diet confounds the relationship between weight and physical activity. Nevertheless, this issue could jeopardize the identification of the whole system as long as the degree to which individuals care about their diet would be a missing variable in the weight equation. In order to deal with this issue, I use a proxy variable; the best one I can count on is the frequency of visits to fast-food restaurants. Therefore, the decision about food consumption will be represented in this dissertation by a variable that describes the frequency of fast-food meals per week.

Physical Activity Physical activity is directly translated into energy expenditure; it is a biological determinant of weight. In this dissertation, it is modeled as a categorical variable that describes the frequency with which several physical activities[12] were performed the week before the respondent was interviewed. The categories are: no physical activity at all (1), one or two times per week (2), 3 to 4 times per week (3), and five or more times per week (4). From the timing assumptions described previously, decisions about physical activity are simultaneously made with smoking, diet, and fertility decisions; this progression implies that smoking diet and fertility decisions do not contemporaneously affect physical activity. The following equation describes the log odds ratio between category k^A and category 1, which is no physical activity at all.

$$\ln\left[\frac{P\left(A_{it}=k^{A}\right)}{P\left(A_{it}=1\right)}\right]=X_{it}\gamma^{A}+Z_{it}\beta^{A}+C_{it}\delta^{A}+\Omega_{it}\phi^{A}+I_{it}\theta^{A}+\mu_{i}^{A}+\upsilon_{it}^{A} \tag{5.3}$$

[12] Walking is not included in the set of physical activities used in this research

$$Where\ \Omega_{it} = [W_{it},\ A_{it-1},\ S_{it-1},\ n_{it-1},\ f_{it-1},\ N_{it}];\ k^A = 1, .., 4;\ t = 2, 3, 4$$

Matrix X_{it} includes individual demographic and socioeconomic exogenous characteristics. Matrix Z_{it} includes amenities of the individual's place of residence. C_{it} is the matrix of career dummies, and Ω_{it} is the vector of state and predetermined variables. The equation also includes a matrix of instruments $I_{it} = [z^A, z^S, z^n, z^f]$, which is composed by exogenous variables that impact the simultaneous behaviors. The error structure is the same as in previous equations in that it allows for unobserved constant and time- varying unobserved heterogeneity.

Smoking Because weight status is the outcome of interest, smoking is treated as an endogenous variable in this study. It is a choice that may be used as a strategy for weight loss or maintenance. It is represented by the categorical variable current smoker or not. The following equation represents the log of the probability ratio between smoking and not smoking:

$$\ln\left[\frac{P\left(S_{it}=1\right)}{P\left(S_{it}=0\right)}\right] = X_{it}\gamma^s + Z_{it}\beta^s + C_{it}\delta^s + \Omega_{it}\phi^s + I_{it}\theta^s + \mu_i^s + \upsilon_{it}^s \tag{5.4}$$

$$Where\ \Omega_{it} = [W_{it},\ A_{it-1},\ S_{it-1},\ n_{it-1},,\ f_{it-1}\ N_{it}];\ t = 2, 3, 4$$

Matrix X_{it} includes individual demographic and socioeconomic exogenous characteristics. Z_{it} includes amenities of the individual's place of residence. C_{it} is the matrix of career dummies, and Ω_{it} is the vector of state and predetermined variables. The equation includes the instruments matrix I_i. The error structure is the same as previous equations.

Childbearing Childbearing is another individual choice variable that can affect health outcomes for women. As a biological process, maternal body size increases during pregnancy. Because weight status can change as a result of weight retention after delivery, this equation is included only in the estimation for women. In order to reduce the complexity of the model, I use a logit model that specifies whether the women had at least one child during period t. The following equation represents the log of the probability ratio between a positive number of deliveries and the reference category (zero new pregnancies).

$$\ln\left[\frac{P\left(n_{it}=1\right)}{P\left(n_{it}=0\right)}\right] = X_{it}\gamma^n + Z_{it}\beta^n + C_{it}\delta^n + \Omega_{it}\phi^n + I_{it}\theta^n + \mu_i^n + \upsilon_{it}^n \tag{5.5}$$

$$Where\ \Omega_{it} = [W_{it},\ A_{it-1},\ S_{it-1},\ n_{it-1},,\ f_{it-1}\ N_{it}];\ t = 2, 3, 4;\ I_{it} = [z^A, z^S, z^n, z^f]$$

The definition of the matrices is the same as in the two previous equations. The error structure is also the same as in previous equations.

Proxy for an Individual's Diet As previously mentioned, the lack of information about the composition of an individual's diet required me to use a proxy variable that describes this individual choice. As a proxy variable, I chose the frequency of meals from fast-food restaurants. A high frequency of fast-food meals would be consistent with a high amount of food consumed, because fast-food restaurants usually serve large portions. In addition, following most of the literature on obesity, one could consider that high consumption of fast food is a signal of poor-quality diet; this is because fast food is usually cheap, high calorie, and calorie-dense. The positive significant relationship between weight and fast food has been noted previously in the literature (Chou, Grossman, & Saffer, 2004). The following equation explains the number of fast food restaurant meals consumed by the respondent during one week. Like all of the previous equations it is a function of matrices X, Z, C, Ω, defined as before. This equation also includes the matrix of instruments I.

$$F_{it} = X_{it}\gamma^F_{kf} + Z_{it}\beta^F_{kf} + C_{it}\delta^F_{kf} + \Omega_{it}\phi^F_{kf} + I_{it}\theta^F_{kf} + \mu^F_i + \upsilon^F_{it} + u_{it} \tag{5.6}$$

$$Where\ \Omega_{it} = [W_{it},\ A_{it-1},\ S_{it-1},\ n_{it-1},,\ f_{it-1}\ N_{it}]; t = 2,3,4;\ I_{it} = [z^A, z^S, z^n, z^f]$$

Weight Status The final equation in the empirical model is the weight status equation. It is modeled as a health outcome produced at the end of each period from the inputs chosen by the individual during the period. The four simultaneous within the period behaviors (food consumption, physical activity, smoking, and fertility) affect the weight produced at the end of the period. It is assumed that neighborhood amenities affect the determination of the weight through their effect on physical activity or other lifestyle simultaneous behaviors. On the other hand, the career-related decisions are assumed to have a direct effect on the determination of the weight. Finally, the weight at the end of the previous period as well as the remaining variables included in vector Ω_{it} are also assumed to determine the weight status. In addition, in this study weight is assumed to be a function of exogenous covariates. The following equation represents the log odds that the individual is overweight at time t

$$\begin{aligned} \ln\left[\frac{P\left(W_{it+1}=1\right)}{P\left(W_{it+1}=0\right)}\right] &= X_{it}\gamma^W + C_{it}\delta^W + W_{it}\phi^W + A_{it}\lambda^W + S_{it}\sigma^W + n_{it}\eta^W \\ &\quad + F_{it}\alpha^W + \mu^W_i + \upsilon^W_{it} \\ where \quad t &= 2,3,4 \end{aligned} \tag{5.7}$$

X_{it}^W includes exogenous characteristics that might increase the probability of obesity. Ω_{it} includes the state and predetermined variables. C_{it} represent the career dummies. A_{it} represents the endogenous contemporaneous physical activity. S_{it} represents the endogenous contemporaneous smoking decision. n_{it} represents the endogenous contemporaneous fertility decision. As in the previous equations μ_i^w represents unobserved individual level characteristics that are constant over time, and v_{it}^w represents time varying unobserved individual characteristics.

4.5 Initial Conditions and Identification Issues

4.5.1 Sources of Identification

One of the advantages of nonlinear systems of dynamic equations is that the identification of the system comes from several sources. Bhargava (1991) shows that in the case of linear dynamic systems under fairly weak conditions the system is identified. The general idea behind identification is based on standard arguments of the dynamic-panel estimation literature (Bhargava and Sargan, 1983; Arellano and Bond, 1991; Mroz and Savage 2006). In dynamic systems, each lagged exogenous variable serves as instrument for the identification of the system. This is because every lag of an exogenous variable could have a separate effect on the contemporaneous value of an endogenous explanatory variable (Mroz and Savage, 2006). In the case of time varying exogenous variables, the longer the temporal dimension of the panel the greater the number of instruments that lead to the over-identification of the system; this is because the whole history of time varying exogenous variables play as instruments for contemporaneous endogenous explanatory variables. In this research the system of equations is non-linear. This is an additional feature that helps to the identification of the model. As discussed in Mroz and Savage (2006), the effect of potential instruments on endogenous contemporary variables depends on the functional form that determines the evolution of any time-varying exogenous variable. This is because the dynamic nature of the system implies that lagged exogenous variables are modified by their previous lags and previous lags of other exogenous variables as well. Finally, the way in which the unobserved heterogeneity is modelled in this research may contribute to identification as well. Conditional on the unobserved heterogeneity components of the composite errors of each equation, the lag of endogenous variables may serve as instruments as well if there is no additional auto-correlation in the remaining iid error components (Yang, Gilleskie and Norton, 2009).

In addition, I incorporate exclusion restrictions in the equations for lifestyle decisions. These are neighborhood amenities and local prices that determine decisions on the practice of physical activity and other lifestyle decisions, but presumably they do not have a direct impact on weight status. Some examples of the variables that I use in this category are: density of different types of recreational facilities, number of parks, and total area of parks available within some radius of the centroid of the individual's

census tract, local prices for cigarettes, junk food, and healthy food. All of these variables are assumed to have a direct impact on physical activity but an indirect impact on weight status. I exclude some individual characteristics such as previous contraception usage and high-school GPA from the obesity equation, I include them in other equations in the system as exclussion restriction as well. These are factors that explain individual's decisions such as fertility and career, but conditional on unobserved heterogeneity I assume they have no direct impact in the weight determination process. Using a log-likelihood ratio test I conclude that all elements excluded from the weight status equation are not jointly significant.

Finally, to ensure identification of the effect of residence amenities upon physical activity and other lifestyle decisions, I use as exclusion restrictions elements that are unique to the residential decision and do not determine directly physical activity practices or other lyfestyle choices. Exogenous variables unique to the residential choice equations may be other characteristics of the residential location (e.g., cost of housing, supply of nearby cultural activities, density of colleges within a specified radius of the respondent's residence, etc.). All of these factors are assumed to influence residence location, but they do not have direct effect on physical activity or other lifestyle decisions.

4.5.2 Initial Conditions

The empirical model estimated in this dissertation is dynamic, which evokes another standard concern with this type of model: the initial condition problem for the lagged endogenous variables. These initial conditions are required because initial values of weight status, the smoking decision, physical activity, food consumption, and the fertility decision cannot be estimated using the specifications described in the previous section. This is because there are no lagged values for the endogenous choices made before the initial period, and means that equations 5.3 to 5.6 cannot be used at the initial period. In order to deal with this situation, in addition to the dynamic equations presented in previous section, I have included in the estimation several reduced-form equations that explain the initial values of the variables described above.

Equations 5.3 to 5.6 are specified for time periods 2, 3, and 4. The initial values equations for those variables at Period One are specified in a similar fashion, but observed right-side variables are strictly exogenous individual characteristics, family background characteristics, original individual high school characteristics, and original individual neighborhood amenities. Unfortunately, during the initial period (Wave I), the respondents were not weighed and measured by the interviewer as was done in the subsequent waves. Nevertheless, for the initial period there are self-reported measures of weight and height. The self-reported information is less than ideal, but it does not cause major estimation problems because it is a dependent variable that is assumed to be measured with error in any case.

4.5.3 Estimation

The methodology for the estimation of all equations in section 5, including the initial condition equations, is based on full information maximum likelihood methods (FIML). A important feature of the empirical model is that it allows for unobserved heterogeneity. As previously mentioned, the specification of each equation in the system includes two unobserved heterogeneity terms (μ_i, υ_{ti}) one time invariant and one time varying. Estimation of the system by FIML typically requires assumptions about the distribution of unobservables μ_i and υ_{ti}; usually, researchers assume multivariate normality.

In this study I propose a more flexible method that does not require any assumption about the distribution of the unobservables. The discrete factor method (DFM) is an extension of the Heckman and Singer (1984) method that approximates the joint distribution of the unobserved heterogeneity as a discrete probability distribution function with a finite number of support points. The probabilities for each support point are jointly estimated with the other parameters in the model (Angeles et al., 1998). The basic idea of semi-parametric methods such as DFM specifies a likelihood function that is conditional upon the values of unobserved heterogeneity (Mroz, 1999) and allows integration over the distribution of the unobserved factors. Using MonteCarlo experiments, it has been proven that when the real underlying distribution for unobserved heterogeneity is normal, the DFM performs very similarly to the models that assume normality. Still, when the real distribution is not normal, the DFM outpaces standard methods in terms of the precision and accuracy of the estimators (Mroz, 1999; Mroz & Guilkey, 1992). More details about the specific implementation of the DFM in this study are provided in the next subsection.

4.5.4 Likelihood Function

In accordance with the DFM, in this study I assume that the cumulative distribution function of the unobserved heterogeneity can be approximated by a step function (Mroz, 1999). Therefore, the discrete distribution for the individual heterogeneity component μ_i is represented by the following expression:

$$\begin{aligned} \Pr\left(\mu_i^e = \boldsymbol{\mu}_{k^e}^e\right) &= \pi\left(q_1\right) \\ \forall\ e &\in \{C, R, A, s, n, F, W\} \\ \forall\ k &\in \{1, ..., K^e\} \\ \text{where } \pi_k &> 0 \text{ and } \sum_{k=1}^{K} \pi_k = 1 \end{aligned}$$

Similarly, the discrete distribution for the time varying unobserved heterogeneity υ_{it} is represented by the following expression:

$$\begin{aligned} \Pr\left(\upsilon_{it} = \upsilon^{e}_{k^e}\right) &= \psi\left(q_2\right) \\ \forall\ e &\in \{C, R, A, s, n, F, W\} \\ \forall\ k &\in \{1, ..., K^e\} \\ \text{where } \psi_l &> 0 \text{ and } \sum_{l=1}^{L} \psi_l = 1 \end{aligned}$$

Where q_1 is the number of mass points allowed for the distribution of the time permanent unobserved heterogeneity, and q_2 is the number of mass points allowed for the distribution of the time permanent unobserved heterogeneity. The unconditional likelihood function (after integrating out the unobserved heterogeneity) for the joint estimation of the system of equations is:

$L\left(\Theta\right) =$

$$\prod_{i=1}^{N}\left\{\sum_{k=1}^{K}\pi_k\left[\prod_{t=2}^{4}\sum_{l=1}^{L}\psi_l\begin{pmatrix} \prod_{k^c=1}^{K^c}\Pr\left(C_{it}=k^c|\mu_k,\upsilon_{lt}\right)^{1\{C_{it}=k^c\}}.1\left\{t>2\right\}. \\ \prod_{k^r\in D}\Pr\left(R_{it}=k^r|\mu_k,\upsilon_{lt}\right)^{1\{R_{it}=k^r\}}.1\left\{t>2\right\}. \\ \prod_{k^a=1}^{K^a}\Pr\left(A_{it}=k^a|\mu_k,\upsilon_{lt}\right)^{1\{A_{it}=k^a\}}. \\ \prod_{k^n=1}^{K^n}\Pr\left(n_{it}=k^n|\mu_k,\upsilon_{lt}\right)^{1\{n_{it}=k^n\}}. \\ \prod_{k^f=1}^{K^f}\Pr\left(F_{it}=k^f|\mu_k,\upsilon_{lt}\right)^{1\left\{F_{it}=k^f\right\}}. \\ \Pr\left(S_{it}=1|\mu_k,\upsilon_{lt}\right)^{1\{S_{it}=1\}}.\left[1-\Pr\left(S_{it}=1|\mu_k,\upsilon_{lt}\right)\right]^{1\{S_{it}=0\}} \\ \Pr\left(W_{it}=1|\mu_k,\upsilon_{lt}\right)^{1\{W_{it}=1\}}.\left[1-\Pr\left(W_{it}=1|\mu_k,\upsilon_{lt}\right)\right]^{1\{W_{it}=0\}} \end{pmatrix}\right]\right\}$$

The whole set of parameters estimated in the model is:

$$\Theta \equiv [\gamma^{e}_{k^e}, \delta^{e}_{k^e}, \phi^{e}_{k^e}, \lambda^{e}_{k^e}, \sigma^{e}_{k^e}, \eta^{e}_{k^e}, \rho^{e}_{k^e}, \tau^{e}_{k^e}, \mu_k, \upsilon_{lt}]$$

$\forall\ e \in \{C, R, A, S, F, n, W\}$,$\forall\ k \in \{1, 2..., K\}$, $\forall\ l \in \{1, 2..., L\}$ .Using some standard normalizations, it is possible to identify all parameters in Θ.

4.5.5 Missing Values and Attrition

Of the initial sample at the beginning of the AddHealth study, less than 50% of the individuals can be tracked through all waves. This inability raises a concern about possible attrition bias. A big share of the high percentage of attrition in AddHealth is explained by the fact that high school seniors in Wave I were not tracked in Wave II. This portion of the attrition would not represent a serious problem, because it is explained by individual ages, which are an exogenous variable that I was able to control for in the regressions. The attrition from Wave I to waves III and IV could certainly be a problem, however. Therefore, for correction of possible bias, I use a methodology based on inverse probability weighting (Horowitz & Manski, 1998; Moffit, et al., 1999; Wooldridge, 2001).

The inverse probability weighting correction uses the probability of selection into the estimation sample, computed from a standard probability model, to weight the individual contributions to the log likelihood function. The probability of selection is computed using exogenous and endogenous characteristics from Wave I (i.e., the initial wave). As can be inferred from the name of the methodology, the weights are the inverse of the individual's probabilities of selection. The advantage of inverse probability weighting over more traditional methodologies based on Heckman's selection procedure is that it does not require exclusion restrictions to achieve identification.

5 Results

In this section I describe the estimation results of the empirical model proposed in previous sections. In addition, in this section I present the result of some simulation-based experiments using parametric bootstrap methodologies. I begin this section by presenting summary statistics of the data. For a convenient presentation of the results, several tables are presented in different Appendices of this dissertation. The unobserved heterogeneity parameters and the probability weights governing the joint distribution of these parameters are presented in Appendix A. The estimation results for the models of Residential Location and Career Decisions are presented in Appendix B and C, respectively.

5.1 Summary Statistics and Sample Description

In the following subsection I present summary statistics of the variables that are used in the estimations of the lifestyle and the weight status equations. Additional variables that describe the neighborhoods in which Addhealth male and female responders live are presented in Appendix D. Additional details about the construction of some of the characteristics are presented in the Data Appendix. The initial sample of the AddHealth study includes more than 20,745 observations. After merging all waves for constructing the panel of individuals an important fraction of observations is lost. The final sample size of individuals who are observed throughout the study is 10,120 (5,520 women and 4,600 men). After

the lost information is given different sources for the missing values, the estimation sample is reduced to 4,400 women and 3,660 men, observed in four different waves.

5.2 Estimations Results of the Model for Women

5.2.1 Obesity Equation

The health outcome modeled in this dissertation is the probability of being obese. Table 2 includes the results of the estimation of this equation for the female sample. Three different specifications are presented, in three panels. The first panel contains the results of an individual logit in which I did not control for any endogeneity issues and unobserved heterogeneity is not modeled in any way. In the other specifications, the obesity equation is jointly estimated together with all other equations in the system and with initial conditions. In the second specification, I use random sampling of the choice set of neighborhoods for the estimation of the residential location equation (RLE). In the third specification, I use an aggregation of the categories into neighborhood types for the estimation of the residential location equation. Both obesity equations in the second and third panels include four permanent and three time-varying unobserved heterogeneity parameters. Inclusion of additional unobserved heterogeneity parameters (permanent and time-varying) does not improve the performance of the models in terms of a significant reduction of the log likelihood function.

One remarkable feature of the estimation of the weight status equation is the persistence of obesity. Being obese is a very inertial state, from which it is difficult to escape. In the simple logit estimation, previous obesity increased by 67 percentage points the probability of being obese in the present. The contribution of 66 percentage points in the jointly estimated models, although a bit lower, is still quite high. Given this level of persistence, the importance of child obesity prevention is crucial. The most important determinant of obesity in the future, by far, is an unhealthy BMI today. Individuals who become obese when they are children or teenagers probably will carry the burden of obesity throughout their lives.

Physical activity is a significant factor in the reduction of the probability of being obese, but only for women who perform physical activity at the highest level. This is the case in the independent logit specification and in the jointly estimated system. In such cases there is a significant reduction of more than 4 percentage points in the probability of obesity. The magnitude of the coefficient for the highest physical activity level is higher in the individual logit estimation. After I control for endogeneity of these variables, there was a reduction in the magnitude of this coefficient, nevertheless it remained highly significant. Clearly, physical activity could be a very important tool with which to tackle the problem of obesity, especially for women. Simulations derived from the estimated model show that

Table 1: Summary Statistics

Variable	Women			Men		
	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
Obese	13203	0.227	0.419	10989	0.217	0.412
At least one Childbirth in the period	13203	0.270	0.444	10989	--	--
Smoker	13203	0.310	0.463	10989	0.363	0.481
PA 1 or 2 times per week	13203	0.152	0.359	10989	0.136	0.342
PA 3 or 4 times per week	13203	0.217	0.413	10989	0.231	0.422
PA 5+ times per week	13203	0.216	0.412	10989	0.359	0.480
Number of Fast food meal/week	13203	2.132	2.008	10989	2.553	2.266
College Student	13203	0.115	0.319	10989	0.091	0.288
High Education/ White Collars	13203	0.124	0.330	10989	0.095	0.294
High Education/ Blue Collars	13203	0.122	0.327	10989	0.153	0.360
Med-Low Education/White Collars	13203	0.111	0.315	10989	0.092	0.288
Med-Low Education/Blue Collars	13203	0.097	0.296	10989	0.169	0.375
Med-Low Education/Not working	13203	0.097	0.296	10989	0.067	0.249
Age	13203	21.855	5.146	10989	22.072	5.177
African American	13203	0.222	0.416	10989	0.177	0.382
Asian	13203	0.050	0.218	10989	0.061	0.240
Hispanic	13203	0.132	0.338	10989	0.144	0.351
1st generation immigrant	13203	0.042	0.201	10989	0.045	0.207
2nd generation immigrant	13203	0.056	0.231	10989	0.065	0.246
Married	13203	0.214	0.410	10989	0.163	0.369
Cohabitating	13203	0.127	0.333	10989	0.119	0.324
Divorced or separated	13203	0.085	0.278	10989	0.079	0.269
Living with Parents	13203	0.505	0.500	10989	0.552	0.497
No. Children younger than 6	13203	0.364	0.694	10989	0.196	0.546
No. Children older than 6	13203	0.129	0.448	10989	0.055	0.310
Family size	13203	3.741	1.778	10989	3.601	1.669
Initial H/H: Step parents	13203	0.141	0.348	10989	0.143	0.350
Initial H/H: Single Father	13203	0.018	0.134	10989	0.029	0.168
Initial H/H: Step Mother	13203	0.199	0.399	10989	0.169	0.375
Initial H/H: Non Parents	13203	0.054	0.226	10989	0.038	0.192
Parents Education: High School	13203	0.249	0.432	10989	0.228	0.420
Parents Education: Some College	13203	0.264	0.441	10989	0.277	0.448
Parents Education: Bachelor	13203	0.178	0.382	10989	0.198	0.399
Parents Education +Bachelor	13203	0.142	0.349	10989	0.152	0.359
Parents Education Missing	13203	0.057	0.231	10989	0.048	0.214
Ground Transportation Terminals by County	13203	0.047	0.305	10989	0.040	0.258
Square miles of parks within 1km of Tract boundaries	13203	0.319	1.110	10989	0.338	1.216
Beta Street connectivity index within 5km Buffers	13203	1.429	0.124	10989	1.423	0.123
Parks within 3km Buffers	13203	5.951	6.962	10989	5.903	7.187
Public PA related amenities within 5km buffers	13203	3.884	5.822	10989	3.783	5.457
Fee required PA related Amenities 5km buffers	13203	6.611	8.151	10989	6.596	7.507
Non PA related Amenities 5km buffers	13203	3.056	9.474	10989	2.881	8.045
Tens of Instruction PA related amenities 5km buffers	13203	12.054	20.657	10989	11.874	18.630
Tens of Membership required PA related amenities 5km buffers	13203	8.442	13.623	10989	8.259	12.215
Tens of Outdoor PA related amenities 5km buffers	13203	5.770	5.706	10989	5.738	5.469
Tens of Amusement Park PA related amenities 5km buffers	13203	0.384	0.790	10989	0.373	0.808
Using any method of contraception (Lagged)	13203	0.855	0.353	10989	0.001	0.033
ACCRA price of a cigarettes Carton, 2005 dollars	13203	33.133	9.047	10989	33.179	9.092
ACCRA Index price for Groceries, 2005 dollars	13203	2.280	0.255	10989	2.281	0.258
ACCRA Index price for Junk food, 2005 dollars	13203	5.333	0.488	10989	5.333	0.490
ACCRA cost of living Index price, 2005 dollars	13203	1.304	0.276	10989	1.306	0.283

Table 2: Weight Status Estimation Results for Women

Variable	[1]:Logit				[2]: Jointly estimated with Random Sampling of the Neighborhood Choice Set[1]				[3]: Jointly estimated with Agregation of the Neighborhood Choice Set[1]			
	Coef	S. D.		Mfx	Coef	S. D.		Mfx	Coef	S. D.		Mfx
Constant	-5.882	1.068	***		-5.849	0.993	***		-6.494	1.001	***	
Obese previous pd	3.807	0.089	***	0.680	3.750	0.096	***	0.671	3.758	0.097	***	0.671
>=1 Childbirths	0.095	0.112		0.009	0.055	0.119		0.006	0.073	0.119		0.007
Smoker	0.012	0.067		0.001	0.000	0.074		0.000	0.012	0.079		0.001
PA 1 or 2 times per week	-0.010	0.085		-0.001	0.014	0.093		0.001	0.020	0.099		0.002
PA 3 or 4 times per week	-0.069	0.086		-0.007	-0.049	0.097		-0.005	-0.047	0.110		-0.005
PA 5+ times per week	-0.452	0.097	***	-0.041	-0.434	0.108	***	-0.040	-0.430	0.119	***	-0.040
# Fast Food meals	-0.009	0.014		-0.001	-0.006	0.031		-0.001	-0.003	0.043		0.000
Students	0.001	0.126		0.000	0.050	0.228		0.005	0.037	0.288		0.004
High Education/ Blue Collars	0.034	0.114		0.003	0.134	0.227		0.014	0.114	0.262		0.011
Med-Low Education/White Collars	0.448	0.115	***	0.048	0.520	0.231	**	0.057	0.491	0.277	*	0.053
Med-Low Education/Blue Collars	0.393	0.127	***	0.041	0.661	0.318	**	0.075	0.607	0.326	*	0.067
Med-Low Education/Not working	0.367	0.128	***	0.039	0.586	0.293	**	0.066	0.525	0.315	*	0.058
Age	0.263	0.097	***	0.032	0.264	0.091	***	0.033	0.263	0.105	***	0.032
Age^2	-0.005	0.002	***	-0.001	-0.005	0.002	***	-0.001	-0.005	0.002	**	-0.001
African American	0.489	0.077	***	0.051	0.508	0.084	***	0.055	0.514	0.094	***	0.055
Asian	-0.117	0.175		-0.011	-0.105	0.338		-0.010	-0.089	0.452		-0.009
Hispanic	0.263	0.102	***	0.027	0.297	0.164	*	0.031	0.312	0.193		0.033
1st generation immigrant	-0.413	0.189	**	-0.037	-0.401	0.368		-0.037	-0.404	0.519		-0.037
2nd generation immigrant	-0.360	0.155	***	-0.033	-0.345	0.227		-0.032	-0.352	0.329		-0.032
Married	0.159	0.095	*	0.016	0.148	0.098		0.015	0.151	0.102		0.015
Cohabitating	0.086	0.098		0.009	0.082	0.103		0.008	0.085	0.104		0.009
Divorced or separated	0.023	0.104		0.002	0.012	0.108		0.001	0.016	0.110		0.002
Living with Parents	0.093	0.093		0.009	0.076	0.097		0.008	0.080	0.104		0.008
No. Children younger than 6	0.031	0.072		0.003	0.035	0.074		0.004	0.029	0.074		0.003
No. Children older than 6	-0.121	0.073	*	-0.011	-0.135	0.080	*	-0.013	-0.134	0.083		-0.013
Family size	0.036	0.023		0.004	0.036	0.024		0.004	0.037	0.024		0.004
Initial H/H: Step parents	-0.240	0.092	***	-0.022	-0.259	0.100	***	-0.025	-0.251	0.117	**	-0.024
Initial H/H: Single Father	-0.027	0.216		-0.003	-0.013	0.407		-0.001	-0.013	0.576		-0.001
Initial H/H: Step Mother	0.142	0.080	*	0.014	0.155	0.087	*	0.016	0.156	0.106		0.016
Initial H/H: Non Parents	-0.144	0.150		-0.014	-0.127	0.199		-0.012	-0.136	0.290		-0.013
Parents Education: High School	-0.030	0.106		-0.003	-0.002	0.172		0.000	-0.002	0.247		0.000
Parents Education: Some College	-0.166	0.107		-0.016	-0.131	0.175		-0.013	-0.132	0.257		-0.013
Parents Education: Bachelor	-0.375	0.120	***	-0.035	-0.331	0.195	*	-0.032	-0.335	0.286		-0.031
Parents Education: +Bachelor	-0.434	0.133	***	-0.039	-0.375	0.213	*	-0.035	-0.373	0.312		-0.035
Parents Education: Missing	0.165	0.160		0.017	0.180	0.275		0.019	0.185	0.361		0.019
Dummy third wave	0.639	0.193	***	0.064	0.539	0.344		0.055	0.573	0.408		0.058
Dummy Fourth wave	0.692	0.256	***	0.073	0.633	0.520		0.068	0.661	0.628		0.070

Notes:

*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

[1]The definition of the neighborhood-type clusters is based on a Ward cluster procedure with 5 categories

the female obesity prevalence rate would be greatly reduced with a generalized implementation of this practice. These simulations will be discussed in the next section.

In the individual logit specification, the dummy variable representing fertility has a positive but insignificant effect. This is also the case in the jointly estimated model specification [2] and specification [3]. In the case of the dummy variable for smokers, the variable is not significant. This is the case in the individual logit specification and in the jointly estimated model as well.

It is difficult to predict the direction of the endogeneity bias in the individual logit estimation. The direction would depend on the correlation of unobservable factors that determine weight status and their correlations with the endogenous variables. Still, it would be reasonable to assume that at least for some variables the direction of the bias is such that it magnifies the real effect of the endogenous variable upon the probability of obesity. Physical activity could be a good illustration of this situation. Individuals who are highly concerned about their health are the ones who tend to engage in more physical activity. Some other characteristics of these individuals, observed and unobserved, are associated with a healthy BMI. For example, more active individuals usually have healthy family backgrounds.

The career-reference category for the specification of this weight status equation is the College-educated white-collar workers category. These are full-time/part-time workers with white-collar jobs who are not attending school; their educational attainment is a college degree or more. In comparison with this category, students and highly educated blue collar workers have no significant increment or reduction in the probability of obesity. This is an interesting result in the sense that it provides evidence that conditional on high educational attainments the individual's occupational choice do not make a difference in terms of changes in the probability of obesity. With regard to the other three categories, in both jointly estimated specifications ([2] and [3]), there is a positive effect associated with white/blue collar workers with low educational attainment and not working nor attending to school individuals with low educational attainment, in comparison with the reference category. These effects are significant in both specifications , the significance level is lower in specification [3], however.

Age is an important explanatory factor for obesity; all other factors constant the probability of obesity increases with age, but in a non-linear way, as one can see from the significance of the age quadratic term. Some exogenous variables (e.g. the dummy variable for married, dummy variables for African American and Hispanic) increase significantly the probability of obesity. In the jointly estimated model, only age, Hispanic, and African American remain significant (some of them only in specification [3]). Other exogenous variables (e.g., inmigrant status, high levels of parental education) significantly reduce the probability of obesity in the individual logit estimation, but some of them are no longer significant in the jointly estimated models. Parents' educational attaiment college or more has a negative effect in the probability of obesity, these effects are significant in specification [2].

5.3 Input Equations Estimation Results for Women

In this subsection I present the estimation results of the four endogenous inputs that are contemporaneously included in the female obesity equation: physical activity, smoking, fertility, and frequency of fast-food meals. Table 3 shows the results of the physical activity equation estimation. The results for the other three inputs are presented in table 4. Two specifications are presented in each table. Specification 1 is an independently estimated multinomial logit. In Specification 2, the PA (physical activity) equation is jointly estimated with all other equations in the system; in addition, in Specification 2 I use random sampling of the choice set of neighborhoods for the estimation of the residential location equation.

5.3.1 Physical Activity

The equation for physical activity (PA) is specified as a multinomial model with four categories associated with different intensity levels of PA: no physical activity at all (1), physical activity one or two times per week (2), physical activity 3 to 4 times per week (3), and physical activity five or more times per week (4). The reference category is the first one (no physical activity at all).

Previous obesity reduces the probability of performing intense PA (i.e., at least five times per week). This effect is significant in the jointly estimated model. An effect that also remains significant after controlling for endogeneity is the negative effect of smoking on the probability of performing intense PA. Individuals are more likely to perform PA if they have performed PA in the past, especially at the intense level. After controlling for endogeneity, the probability of intense PA increases by 9 percentage points for individuals who performed medium PA (3 to 4 times per week) in the previous period, and 23 percentage points for those who performed intense PA in the previous period. There is a significant reduction in the probability of medium and intense PA with each fast-food meal reported per week. There is a significant reduction in the probability of intense PA of several careers in comparison with the reference category college-educated white collar workers.

The probability of intense PA decreases with age in a non-linear way; this reduction is significant in all specifications presented in the table 3. Some demographic factors significantly reduce the probability of PA as well; for example the dummy for African American, which is associated with a significant reduction of more than 5 percentage points in the probability of intense levels of PA relative to the reference category (zero physical activity). Having more small children (younger than 6) also reduces the probability of PA. The probability of positive PA levels decreases significantly for cohabitating women, the same it is true for married women, but in this case the effect is not significant in the jointly estimated model.

To summarize, some of the most remarkable features of the estimation are that female respondents in AddHealth are significantly more likely to perform intense PA in the current period if during the previous period they did not smoke but did practice medium or intense PA. In addition, more frequent fast-food meals in the previous period are associated with a significant reduction in the probability of performing medium and intense PA during the current period. This evidence seems to suggest that healthy behaviors in the past increase the probability of healthy behaviors in the present.

This model allows us to test an additional hypothesis: whether neighborhood amenities have a real impact in the determination of obesity. I test this hypothesis by evaluating the effect that a set of neighborhood amenities has on the probability of performing PA. In the individual multinomial logit for PA, several neighborhood amenities have a significant impact upon the probability of PA, especially at the medium and intense levels. Variables such as square miles of community parks, the number of parks in the neighborhood, the "beta" street conectivity index, and the amount of fee required-PA-related facilities in the neighborhood have a positive effect in the probability of at least one non-zero PA category. Square miles of parks and fee-required facilities are significant at 5% level, and the number of parks is almost significant at 10% level. After controlling for endogeneity only one variable remains significant: the square mileage of community parks within one 1 km of a woman's neighborhood. This variable has a positive significant effect upon low and medium PA and is barely significant (based on a 10% level of significance) in the intense PA category. Other variables that represent neighborhood amenities have expected signs, but they are not significant.

Table 3: Physical Activity Estimation Results for Women

Variable	[1]: Logit PA=2 relative to PA=1			[2]: Jointly Estimated PA=2 relative to PA=1			[1]: Logit PA=3 relative to PA=1			[2]: Jointly Estimated PA=3 relative to PA=1			[1]: Logit PA=4 relative to PA=1			[2]: Jointly Estimated PA=4 relative to PA=1		
	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx
Constant	-2.078	1.383		-2.176	0.997 **		0.416	1.296		-0.240	0.952		5.824	1.326 ***		5.180	0.962 ***	
Obese (t-1)	-0.035	0.086	0.006	-0.020	0.094	0.007	-0.095	0.088	0.001	-0.078	0.096	0.002	-0.235	0.100 ***	-0.025	-0.224	0.108 **	-0.025
>=1 Childbirths (t-1)	0.091	0.132	-0.009	0.093	0.212	-0.010	0.210	0.141	0.003	0.212	0.260	0.003	0.410	0.160 ***	0.043	0.418	0.339	0.044
Smoker (t-1)	-0.185	0.069 ***	-0.015	-0.182	0.072 ***	-0.015	-0.071	0.067	0.007	-0.066	0.070	0.008	-0.184	0.071 ***	-0.015	-0.186	0.075 ***	-0.016
PA 1 or 2 times per week (t-1)	0.338	0.096 ***	0.014	0.335	0.125 ***	0.013	0.459	0.100 ***	0.036	0.452	0.130 ***	0.035	0.397	0.119 ***	0.016	0.403	0.172 ***	0.017
PA 3 or 4 times per week (t-1)	0.320	0.093 ***	-0.022	0.312	0.123 ***	-0.023	0.766	0.093 ***	0.041	0.758	0.126 ***	0.042	1.058	0.106 ***	0.094	1.051	0.159 ***	0.094
PA 5+ times per week (t-1)	0.582	0.098 ***	-0.043	0.583	0.131 ***	-0.044	1.138	0.097 ***	0.020	1.143	0.134 ***	0.022	2.056	0.106 ***	0.229	2.058	0.164 ***	0.230
# Fast Food meals (t-1)	-0.021	0.016	0.001	-0.024	0.017	0.001	-0.043	0.016 ***	-0.002	-0.043	0.017 ***	-0.002	-0.069	0.017 ***	-0.006	-0.070	0.018 ***	-0.006
Students	0.139	0.120	0.031	0.127	0.206	0.031	-0.136	0.123	-0.015	-0.163	0.210	-0.018	-0.198	0.131	-0.024	-0.202	0.230	-0.022
High Education/ Blue Collars	-0.166	0.111	-0.002	-0.172	0.200	-0.003	-0.149	0.110	0.011	-0.165	0.194	0.007	-0.476	0.123 ***	-0.049	-0.443	0.215 **	-0.044
Med-Low Education/White Collars	-0.181	0.118	-0.001	-0.182	0.213	-0.002	-0.271	0.121 **	-0.012	-0.285	0.213	-0.016	-0.394	0.133 ***	-0.032	-0.363	0.235	-0.027
Med-Low Education/Blue Collars	-0.272	0.132 **	0.011	-0.299	0.282	0.008	-0.673	0.146 ***	-0.049	-0.734	0.296 ***	-0.059	-0.734	0.158 ***	-0.051	-0.691	0.315 **	-0.042
Med-Low Education/Not working	-0.558	0.140 ***	-0.029	-0.603	0.280 **	-0.031	-0.563	0.144 ***	-0.024	-0.663	0.277 ***	-0.037	-0.733	0.161 ***	-0.048	-0.756	0.306 ***	-0.045
Age	0.232	0.102 **	0.039	0.222	0.122 *	0.039	-0.023	0.096	-0.010	-0.033	0.109	-0.010	-0.537	0.097 ***	-0.048	-0.550	0.112 ***	-0.049
Age^2	-0.004	0.002 **	-0.001	-0.004	0.003 *	-0.001	0.000	0.002	0.000	0.000	0.002	0.000	0.010	0.002 ***	0.001	0.011	0.002 ***	0.001
African American	-0.327	0.084 ***	-0.010	-0.312	0.096 ***	-0.010	-0.368	0.081 ***	-0.009	-0.346	0.094 ***	-0.007	-0.608	0.088 ***	-0.050	-0.593	0.103 ***	-0.050
Asian	0.336	0.158 **	0.058	0.357	0.251	0.060	-0.060	0.161	-0.012	-0.048	0.263	-0.012	-0.199	0.170	-0.035	-0.175	0.284	-0.033
Hispanic	0.105	0.108	0.023	0.117	0.141	0.025	-0.130	0.109	-0.016	-0.123	0.146	-0.016	-0.125	0.114	-0.013	-0.120	0.154	-0.013
1st generation immigrant	-0.488	0.173 ***	-0.031	-0.487	0.297	-0.031	-0.512	0.177 ***	-0.038	-0.508	0.286 *	-0.038	-0.373	0.185 **	-0.006	-0.375	0.311	-0.007
2nd generation immigrant	-0.306	0.150 **	-0.030	-0.306	0.245	-0.030	-0.060	0.144	0.011	-0.062	0.205	0.011	-0.153	0.154	-0.008	-0.156	0.241	-0.009
Married	-0.075	0.098	0.002	-0.078	0.110	0.003	-0.166	0.101 *	-0.012	-0.177	0.119	-0.013	-0.186	0.111 *	-0.013	-0.197	0.136	-0.014
Cohabitating	-0.251	0.102 ***	-0.008	-0.250	0.115 **	-0.008	-0.361	0.105 ***	-0.025	-0.358	0.124 ***	-0.025	-0.358	0.117 ***	-0.020	-0.361	0.142 ***	-0.020
Divorced or separated	-0.103	0.111	-0.006	-0.093	0.120	-0.007	-0.218	0.119 *	-0.031	-0.200	0.134	-0.030	0.019	0.125	0.019	0.044	0.145	0.022
Living with Parents	-0.100	0.095	-0.003	-0.100	0.102	-0.003	-0.173	0.096 *	-0.017	-0.167	0.104	-0.016	-0.108	0.103	-0.001	-0.108	0.112	-0.002
No. Children younger than 6	-0.139	0.056 ***	-0.001	-0.143	0.060 ***	-0.002	-0.225	0.062 ***	-0.014	-0.230	0.067 ***	-0.014	-0.274	0.070 ***	-0.020	-0.281	0.078 ***	-0.020
No. Children older than 6	-0.020	0.093	0.001	-0.021	0.131	0.000	-0.049	0.105	-0.003	-0.036	0.163	-0.001	-0.061	0.117	-0.005	-0.056	0.209	-0.005

Continuing in the next page

Table 3: Physical Activity Estimation Results for Women (Continuatued from Previous Page)

Variable	[1]: Logit PA=2 relative to PA=1			[2]: Jointly Estimated PA=2 relative to PA=1			[1]: Logit PA=3 relative to PA=1			[2]: Jointly Estimated PA=3 relative to PA=1			[1]: Logit PA=4 relative to PA=1			[2]: Jointly Estimated PA=4 relative to PA=1		
	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx
Family size	0.022	0.023	0.003	0.024	0.024	0.004	-0.018	0.023	-0.004	-0.015	0.024	-0.004	0.002	0.024	0.001	0.006	0.025	0.001
Initial H/H: Step parents	-0.117	0.090	0.001	-0.111	0.109	0.001	-0.298	0.089 ***	-0.031	-0.288	0.109 ***	-0.030	-0.174	0.093 *	-0.003	-0.166	0.117	-0.002
Initial H/H: Single Father	-0.554	0.248 **	-0.044	-0.551	0.713	-0.044	-0.538	0.229 ***	-0.053	-0.532	0.567	-0.053	-0.138	0.229	0.029	-0.126	0.562	0.030
Initial H/H: Step Mother	0.067	0.083	0.014	0.069	0.100	0.015	-0.152	0.082 *	-0.026	-0.148	0.099	-0.025	-0.003	0.087	0.007	0.001	0.108	0.007
Initial H/H: Non Parents	-0.233	0.157	-0.008	-0.226	0.332	-0.008	-0.313	0.158 **	-0.019	-0.302	0.354	-0.019	-0.337	0.175 *	-0.020	-0.324	0.418	-0.019
Parents Education: High School	-0.178	0.112	-0.015	-0.182	0.251	-0.015	-0.034	0.114	0.013	-0.044	0.218	0.012	-0.195	0.122	-0.019	-0.204	0.261	-0.020
Parents Education: Some College	-0.067	0.112	-0.015	-0.070	0.257	-0.015	0.099	0.114	0.013	0.089	0.224	0.012	0.087	0.120	0.008	0.078	0.263	0.008
Parents Education: Bachelor	0.074	0.121	-0.004	0.070	0.276	-0.003	0.158	0.122	0.007	0.138	0.241	0.005	0.237	0.128 *	0.021	0.225	0.279	0.021
Parents Education: +Bachelor	0.099	0.131	-0.010	0.091	0.303	-0.010	0.301	0.130 **	0.021	0.274	0.259	0.018	0.371	0.136 ***	0.031	0.350	0.300	0.030
Parents Education: Missing	-0.024	0.168	0.000	-0.031	0.391	0.000	-0.024	0.174	0.001	-0.035	0.351	0.000	-0.062	0.190	-0.006	-0.075	0.430	-0.007
Ground Transportation Terminals by County/ Area	0.141	0.166	0.023	0.149	0.318	0.022	0.041	0.160	0.010	0.062	0.320	0.012	-0.154	0.194	-0.028	-0.135	0.427	-0.027
Miles^2 of parks within 1km of Tract boundaries	0.051	0.024 **	0.003	0.051	0.025 **	0.003	0.070	0.023 ***	0.007	0.070	0.024 ***	0.007	0.039	0.026	-0.001	0.041	0.027	0.000
Beta Street conectivity index within 5km Buffers	-0.455	0.314	-0.068	-0.452	0.682	-0.068	0.581	0.297 *	0.116	0.556	0.584	0.111	0.041	0.315	-0.026	0.055	0.647	-0.021
Parks within 3km Buffers	0.103	0.065	0.011	0.092	0.089	0.010	0.022	0.062	-0.005	0.019	0.087	-0.004	0.065	0.065	0.004	0.061	0.092	0.004
Public PA related amenities within 5km buffers	-0.013	0.102	0.000	-0.005	0.145	0.000	-0.078	0.100	-0.013	-0.065	0.138	-0.012	0.030	0.106	0.010	0.042	0.146	0.010
Fee required PA related Amenities 5km buffers	0.061	0.120	-0.002	0.065	0.214	-0.002	0.049	0.116	-0.010	0.053	0.218	-0.010	0.246	0.123 **	0.030	0.254	0.242	0.031
Non PA related Amenities 5km buffers	0.008	0.093	0.005	0.013	0.152	0.006	-0.007	0.087	0.005	-0.006	0.151	0.006	-0.109	0.104	-0.015	-0.117	0.185	-0.016
Instruction PA related amenities 5km buffers	-0.081	0.053	-0.013	-0.084	0.060	-0.013	0.033	0.050	0.004	0.030	0.055	0.004	0.062	0.054	0.009	0.055	0.060	0.008
Membership required PA amenities 5km buffers	0.023	0.085	0.010	0.021	0.114	0.010	-0.064	0.081	-0.003	-0.070	0.113	-0.004	-0.132		-0.015	-0.126	0.126	-0.014
Outdoor PA related amenities 5km buffers	0.033	0.122	0.008	0.019	0.246	0.007	0.051	0.119	0.018	0.042	0.254	0.018	-0.183	0.129	-0.029	-0.189	0.286	-0.029
Amusement Park PA related amenities 5km buffers	0.617	0.471	0.080	0.545	0.960	0.075	0.111	0.475	-0.023	0.084	0.971	-0.016	0.238	0.496	0.003	0.099	0.969	-0.012
Using any method of contraception (Lagged)	0.138	0.086	-0.002	0.140	0.088	-0.002	0.239	0.087 ***	0.011	0.242	0.090 ***	0.011	0.351	0.096 ***	0.029	0.355	0.100 ***	0.029
ACCRA price of a cigarrete Carton, 2005 dollars	0.019	0.010 **	0.002	0.019	0.016	0.002	0.004	0.009	-0.001	0.002	0.014	-0.001	0.014	0.010	0.001	0.012	0.014	0.001
ACCRA Index price for Groceries, 2005 dollars	0.406	0.247 *	0.078	0.403	0.592	0.080	-0.294	0.240	-0.051	-0.297	0.551	-0.051	-0.129	0.252	-0.017	-0.157	0.565	-0.021
ACCRA Index price for Junk food, 2005 dollars	-0.036	0.127	-0.008	-0.039	0.232	-0.009	0.076	0.122	0.011	0.077	0.236	0.012	0.034	0.127	0.001	0.035	0.250	0.001
ACCRA cost of living Index price, 2005 dollars	-0.276	0.299	-0.055	-0.317	0.606	-0.058	0.406	0.282	0.050	0.383	0.594	0.050	0.399	0.290	0.039	0.361	0.603	0.036
Dummy third wave	-2.897	0.290 ***	-0.155	-2.824	0.552 ***	-0.160	-2.868	0.279 ***	-0.171	-2.706	0.474 ***	-0.160	-2.568	0.294 ***	-0.099	-2.430	0.500 ***	-0.092
Dummy Fourth wave	-2.429	0.336 ***	-0.169	-2.363	0.684 ***	-0.174	-2.114	0.327 ***	-0.141	-1.958	0.596 ***	-0.130	-1.352	0.347 ***	0.005	-1.225	0.647 *	0.011

*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

5.3.2 Smoking, Childbirth, and Fast Food Meals for Women

Table 4 presents the results of the estimation of the equations for the smoking decision, the fertility decision (at least one childbirth during the period), and the linear model for the number of fast-food meals. As before, two specifications are presented for each equation. Specification 1 is the estimation of each equation separately. Specification 2 represents each equation estimated jointly with all remaining equations of the system.

Female respondents in AddHealth are more likely to smoke regularly during the current period if they smoked during the previous period, if they were obese during the previous period, and if they did not practice any physical activity during the previous last period. After controlling for unobserved heterogeneity, all these relationships remain significant. Again, this seems to suggest that unhealthy practices in the past may explain current unhealthy lifestyles. For previous smokers there is an increment of 49 percentage points in the probability of smoking, whereas for obese in the last period there is an increment of 2 percentage points. Some demographic variables have a negative effect upon the probability of smoking, even after controlling for unobserved heterogeneity. The dummy variables for African American and Hispanic are negative, which implies a reduction in the probability of smoking in comparison with the reference category (white females). First and second generation immigrant status have also negative effects in the probability of smoking, but they are not significant in the jointly estimated model. Other demographic variables have expected signs but they are not significant after controlling for unobserved heterogeneity.

Table 4: Smoking, Pregnancies, and Fast Food Meals Estimation Results

Variable	Smoking: Logit [1]			Smoking: Jointly Estimated [2]			Childbirth: Logit [1]			Childbirth: Jointly Estimated [2]			Fast Food Meals: OLS [1]			Fast Food Meals: Jointly		
	Coef	S. D.	Mfx	Coef	S. D.	Mfx	Coef	S. D.	Mfx	Coef	S. D.	Mfx	Coef	S. D.		Coef	S. D.	
Constant	-1.560	0.983		-1.314	0.992		-3.509	3.349		-3.896	3.258		1.943	0.650	***	6.129	0.897	***
Obese (t-1)	0.143	0.075	* 0.020	0.141	0.085	* 0.020	-0.129	0.175	-0.003	-0.234	0.300	-0.005	-0.033	0.051		-0.013	0.043	
>=1 Childbirths (t-1)	0.284	0.117	*** 0.041	0.282	0.122	** 0.041	-3.664	0.231	*** -0.083	-3.688	0.436	*** -0.082	0.178	0.077	**	0.148	0.069	**
Smoker (t-1)	2.475	0.053	*** 0.491	2.475	0.054	*** 0.490	0.116	0.147	0.003	0.101	0.205	0.002	-0.042	0.038		-0.051	0.033	
PA 1 or 2 times per week (t-1)	-0.192	0.090	** -0.027	-0.189	0.102	* -0.026	0.202	0.209	0.005	0.227	0.362	0.005	-0.096	0.058	*	-0.047	0.049	
PA 3 or 4 times per week (t-1)	-0.129	0.084	-0.018	-0.124	0.095	-0.017	0.141	0.200	0.003	0.154	0.351	0.003	-0.143	0.054	***	-0.058	0.046	
PA 5+ times per week (t-1)	-0.149	0.086	* -0.021	-0.147	0.095	-0.021	0.167	0.210	0.004	0.175	0.369	0.004	-0.073	0.055		-0.011	0.047	
# Fast Food meals (t-1)	0.033	0.014	*** 0.005	0.033	0.015	** 0.005	0.049	0.034	0.001	0.048	0.037	0.001	0.475	0.009	***	0.477	0.009	***
Students	0.280	0.123	** 0.041	0.284	0.245	0.041	0.386	0.293	0.009	0.505	0.608	0.012	0.105	0.074		0.103	0.125	
High Education/ Blue Collars	0.484	0.111	*** 0.071	0.480	0.238	** 0.071	0.419	0.249	* 0.010	0.609	0.400	0.014	0.363	0.067	***	0.148	0.105	
Med-Low Education/White Collars	0.592	0.115	*** 0.089	0.589	0.254	*** 0.088	0.622	0.249	*** 0.015	0.810	0.410	** 0.020	0.474	0.072	***	0.290	0.116	***
Med-Low Education/Blue Collars	0.814	0.124	*** 0.126	0.816	0.325	*** 0.126	0.849	0.270	*** 0.021	1.315	0.488	*** 0.036	0.688	0.079	***	0.353	0.143	***
Med-Low Education/Not working	0.933	0.126	*** 0.145	0.953	0.305	*** 0.149	1.846	0.273	*** 0.061	2.241	0.474	*** 0.081	0.061	0.080		0.132	0.141	
Age	0.052	0.070	0.008	0.052	0.111	0.008	0.151	0.241	0.004	0.175	0.299	0.005	0.049	0.046		0.057	0.053	
Age2	-0.002	0.002	0.000	-0.002	0.002	0.000	-0.004	0.005	0.000	-0.005	0.006	0.000	-0.001	0.001		-0.001	0.001	
African American	-0.833	0.075	*** -0.114	-0.841	0.082	*** -0.115	0.952	0.174	*** 0.023	0.979	0.346	*** 0.023	0.419	0.047	***	0.296	0.042	***
Asian	-0.041	0.148	-0.006	-0.044	0.382	-0.006	-0.083	0.415	-0.002	-0.093	0.717	-0.002	0.233	0.092	***	0.186	0.077	***
Hispanic	-0.378	0.094	*** -0.052	-0.379	0.175	** -0.052	0.128	0.243	0.003	0.163	0.504	0.004	0.194	0.061	***	0.188	0.052	***
1st generation immigrant	-0.823	0.190	*** -0.106	-0.826	0.609	-0.106	0.094	0.431	0.002	0.102	0.713	0.002	0.045	0.099		0.020	0.085	
2nd generation immigrant	-0.394	0.140	*** -0.053	-0.395	0.278	-0.053	0.589	0.341	* 0.014	0.609	0.670	0.015	-0.109	0.084		-0.083	0.070	
Married	-0.486	0.092	*** -0.066	-0.487	0.094	*** -0.066	0.954	0.194	*** 0.025	0.949	0.326	*** 0.024	-0.097	0.058	*	-0.045	0.051	
Cohabitating	0.156	0.090	* 0.022	0.153	0.093	* 0.022	0.692	0.197	*** 0.017	0.680	0.327	** 0.016	0.013	0.060		0.019	0.053	
Divorced or separated	0.735	0.097	*** 0.112	0.728	0.099	*** 0.111	0.699	0.210	*** 0.017	0.669	0.404	* 0.016	0.020	0.065		-0.107	0.059	*
Living with Parents	-0.162	0.086	* -0.023	-0.166	0.089	* -0.023	-0.384	0.204	* -0.008	-0.416	0.411	-0.009	0.246	0.055	***	0.213	0.051	***
No. Children younger than 6	-0.190	0.050	*** -0.026	-0.187	0.054	*** -0.026	6.975	0.185	*** 0.799	7.046	0.326	*** 0.799	-0.083	0.033	***	-0.035	0.028	
No. Children older than 6	-0.034	0.082	-0.005	-0.037	0.087	-0.005	1.930	0.151	*** 0.082	1.928	0.243	*** 0.079	-0.003	0.055		-0.045	0.049	
Family size	0.061	0.019	*** 0.009	0.061	0.020	*** 0.009	0.015	0.052	0.000	0.011	0.076	0.000	0.001	0.013		-0.011	0.010	

Continuing in the next page

Table 4: Smoking, Pregnancies, and Fast Food Meals Estimation Results (Continuation)

Variable	Smoking: Logit [1]			Smoking: Jointly Estimated [2]			Childbirth: Logit [1]			Childbirth: Jointly Estimated [2]			Fast Food Meals: OLS [1]			Fast Food Meals: Jointly		
	Coef	S. D.		Mfx	Coef	S. D.	Mfx	Coef	S. D.	Mfx	Coef	S. D.	Mfx	Coef	S. D.		Coef	S. D.	
Initial H/H: Step parents	0.281	0.073	***	0.041	0.280	0.088	*** 0.041	0.234	0.193	0.005	0.200	0.357	0.004	0.015	0.050		-0.005	0.043	
Initial H/H: Single Father	0.113	0.185		0.016	0.111	0.545	0.016	0.150	0.494	0.003	0.152	0.836	0.003	0.059	0.124		0.002	0.112	
Initial H/H: Step Mother	0.231	0.070	***	0.033	0.228	0.088	*** 0.033	0.249	0.176	0.006	0.243	0.310	0.005	0.044	0.047		0.012	0.039	
Initial H/H: Non Parents	0.196	0.135		0.028	0.191	0.203	0.027	0.365	0.296	0.008	0.370	0.509	0.008	0.104	0.088		0.012	0.077	
Parents Education: High School	-0.058	0.096		-0.008	-0.057	0.272	-0.008	-0.289	0.218	-0.006	-0.243	0.330	-0.005	0.018	0.063		0.043	0.054	
Parents Education: Some College	-0.047	0.096		-0.007	-0.044	0.278	-0.006	-0.325	0.222	-0.007	-0.254	0.343	-0.005	0.003	0.063		0.038	0.054	
Parents Education: Bachelor	-0.193	0.106	*	-0.027	-0.190	0.304	-0.026	-0.726	0.258	*** -0.015	-0.622	0.424	-0.013	-0.097	0.068		-0.047	0.060	
Parents Education: +Bachelor	-0.222	0.114	*	-0.031	-0.215	0.321	-0.030	-0.915	0.305	*** -0.019	-0.782	0.612	-0.016	-0.252	0.073	***	-0.166	0.064	***
Parents Education: Missing	-0.075	0.146		-0.010	-0.072	0.419	-0.010	-0.985	0.340	*** -0.020	-0.966	0.571	* -0.019	-0.105	0.096		-0.049	0.083	
Ground Transportation Terminals by County/ Area	0.115	0.151		0.017	0.109	0.161	0.016	0.235	0.472	0.005	0.230	0.818	0.005	0.146	0.100		0.070	0.082	
Square miles of parks within 1km of Tract boundaries	0.009	0.023		0.001	0.009	0.024	0.001	0.014	0.059	0.000	0.016	0.062	0.000	0.000	0.014		-0.003	0.012	
Beta Street conectivity index within 5km Buffers	0.177	0.252		0.026	0.195	0.675	0.028	-1.387	0.760	* -0.028	-1.300	0.840	-0.026	-0.126	0.167		-0.058	0.238	
Parks within 3km Buffers	-0.179	0.057	***	-0.025	-0.179	0.071	*** -0.025	0.092	0.161	0.002	0.078	0.320	0.002	-0.052	0.036		-0.064	0.030	**
Public PA related amenities within 5km buffers	0.003	0.098		0.000	-0.002	0.127	0.000	0.184	0.276	0.004	0.169	0.665	0.004	0.210	0.059	***	0.135	0.076	*
Fee required PA related Amenities 5km buffers	0.110	0.105		0.016	0.111	0.117	0.016	0.027	0.271	0.001	-0.004	0.588	0.000	0.052	0.067		0.052	0.058	
Non PA related Amenities 5km buffers	0.046	0.083		0.007	0.045	0.093	0.006	-0.198	0.276	-0.004	-0.178	0.548	-0.004	-0.010	0.054		0.015	0.050	
Instruction PA related amenities 5km buffers	-0.065	0.047		-0.009	-0.065	0.053	-0.009	0.028	0.128	0.001	0.042	0.248	0.001	-0.072	0.030	***	-0.065	0.029	**
Membership required PA amenities 5km buffers	0.073	0.075		0.010	0.077	0.085	0.011	-0.261	0.206	-0.006	-0.278	0.410	-0.006	-0.018	0.048		-0.007	0.042	
Outdoor PA related amenities 5km buffers	-0.102	0.110		-0.014	-0.099	0.123	-0.014	0.232	0.288	0.005	0.254	0.619	0.006	0.049	0.070		0.036	0.071	
Amusement Park PA related amenities 5km buffers	-0.229	0.455		-0.031	-0.227	0.914	-0.031	1.310	1.074	0.039	1.289	0.999	0.037	-0.710	0.275	***	-0.348	0.759	
Using any method of contraception (Lagged)	-0.141	0.073	*	-0.020	-0.143	0.076	* -0.020	-0.193	0.166	-0.004	-0.193	0.226	-0.004	-0.015	0.049		-0.043	0.043	
ACCRA price of a cigarrete Carton, 2005 dollars	-0.016	0.008	**	-0.002	-0.015	0.012	-0.002	0.010	0.022	0.000	0.006	0.033	0.000	-0.012	0.005	***	-0.002	0.009	
ACCRA Index price for Groceries, 2005 dollars	0.410	0.205	**	0.061	0.403	0.458	0.060	-0.701	0.576	-0.014	-0.697	0.786	-0.014	-0.470	0.136	***	-0.399	0.186	**
ACCRA Index price for Junk food, 2005 dollars	-0.106	0.101		-0.015	-0.105	0.134	-0.015	0.111	0.301	0.003	0.155	0.481	0.003	0.003	0.066		-0.016	0.072	
ACCRA cost of living Index price, 2005 dollars	-0.080	0.216		-0.011	-0.080	0.461	-0.011	-0.317	0.709	-0.007	-0.244	0.855	-0.005	0.027	0.143		0.020	0.246	
Dummy third wave	-0.718	0.230	***	-0.097	-0.763	0.334	** -0.103	0.423	0.651	0.009	0.140	0.719	0.003	0.317	0.150	**	-0.361	0.427	
Dummy Fourth wave	-0.168	0.272		-0.023	-0.212	0.487	-0.030	1.206	0.750	0.031	1.003	0.844	0.025	0.042	0.177		-0.619	0.478	

*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

There is a negative effect for married female respondents in the probability of smoking in comparison with single females, which is significant in the jointly estimated model. There is a positive and significant effect in the smoking probability for females who are divorced or cohabitating. Female respondents are more likely to smoke if they were observed at the beginning of the study in households with at least one step-parent or a single mother, in comparison with females who were observed at the beginning of the study with two biological parents. In addition, the number of children younger than six reduces significantly the probability of smoking for females.

Female AddHealth respondents who were obese in the previous period are less likely to have had at least one child during the current period, but this effect is not significant. For women who had a childbirth in the previous period, there is an significant reduction in the probability of childbirth in the current period (8 percentage points). In comparison with college-esducated white collar workers there is a significant increase in the probability of childbirth for all other careers except by students. These effects remain significant after controlling for unobserved heterogeneity. Married and cohabiting females are more likely to have children in comparison with single women. African Americans females have a higher probability of pregnancy than white females and this effect is significant at 5% confidence level.

Female AddHealth respondents consume fewer fast-food meals per week if they smoked in the previous period, but this effect is not significant in all specifications. Previous consumption of fast-food meals greatly explains current consumption. On average, African American, Hispanic, and Asian females consume more fast-food meals than white females do. Married and divorsed women consume less fast food meals in comparison with single women, but this negative effect is only significant for divorsed women in the jointly estimated model. Female respondents who still live with their parents significantly consume more fast-food meals per week, whereas females whose parents are college-educated consume fewer fast-food meals per week.

5.4 Simulations Using Parametric Bootstrap and Fit of the Model for Women

The model estimated in this dissertation represents a system of different choices and behaviors, all of which have a final direct or indirect effect on the probability of obesity. Marginal effects per period of a variable only partly describe the effect that a marginal change in that variable has on the obesity probability. Given the multidimensional and dynamic nature of the model, changes in one variable might have a direct effect on the obesity equation; however, they may also have many indirect effects through the effect that the original perturbation has on the system of variables that also determine obesity. Such situations make it difficult to measure the ultimate impact of any exogenous or endogenous covariates on the key outcomes, as well as to follow the pathways through which these variables operate. This difficulty can be overcome, however, by using simulations.

The basic strategy the simulations is to use the estimated coefficients, mass points, and probability weights from the reduced-form equations to predict values for the endogenous inputs of the obesity equation (e.g., physical activity, smoking, fertility, and the proxy for food consumption)[13]. This is done by comparing the predicted probability of each behavior with random draws of a standard uniform distribution. Then these predicted values are used, along with the actual observed values of the exogenous variables in the obesity equation, to predict the probability of being obese. By using the same strategy of comparing the estimated probability with a random draw from a uniform distribution and then averaging the simulated outcome across respondents, I was able to get model-predicted obesity prevalence rates. These simulated prevalence obesity rates, among other things, allow for testing the fit of the model by comparing the prediction with the real prevalence observed in the AddHealth sample.

At the end of the process described in the previous paragraph, point estimates of the obesity prevalence rates can be calculated. One can obtain an expectation for the estimate and a confidence interval by using bootstrapping methodologies. In this study I use parametric bootstrapping methods. Parametric bootstrap make use of the following asymptotic result that holds for MLE estimation.

$$\beta \sim N(\hat{\beta}, cov_{\hat{\beta}})$$

In other words, it is assumed that the entire set of estimated coefficients, mass points, and mass-point weights follow a multivariate normal distribution that is centered at the estimated values of the parameters, with a covariance matrix equal to the estimated covariance matrix for the entire set of parameters (Angeles, Guilkey, & Mroz, 2005). I randomly drew 1,000 parameter vectors in order to conduct the simulation exercises and used the standard deviation across the 1,000 bootstrap samples to construct confidence intervals for the predicted obesity rates.

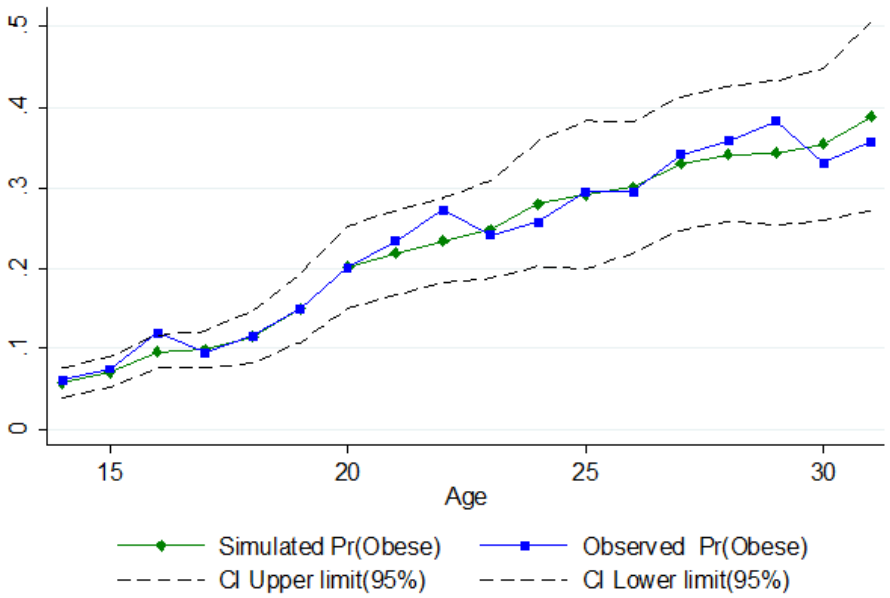
5.4.1 Fit of the Model

Using the procedure described in the introduction of this subsection, I was able to compare the model predictions for obesity prevalence rates with the prevalence rates observed in the data. This comparison provided a straightforward way to see how well the model was predicting the data and how accurate those predictions were. Figure 2 contains the predictions and confidence intervals (at a 95% significance level) of the obesity prevalence rate at all ages at which respondents are observed in the AddHealth study. The green line represents the obesity prevalence rate estimated, averaged through all bootstrap samples, for the jointly estimated model (Specification 2). The blue line represents the obesity prevalence rate, which was computed using the all the respondents in AddHealth observed at any of the specific ages

[13] In order to get a unique prediction of a specific input probability, we compute the expectation of the latent indirect utility level over the distribution of the different individual's types, given the values of the unobserved heterogeneity parameters.

represented in the horizontal axis. The upper and lower limits of the confidence intervals are represented by the black dotted lines. On average, the model predictions of the female obesity prevalence rate are a bit smaller than the observed prevalence for the female AddHealth respondents. Nevertheless, the model captures well the evolution of female obesity prevalence.

Figure 2: Model Predictions of Female Obesity Prevalence

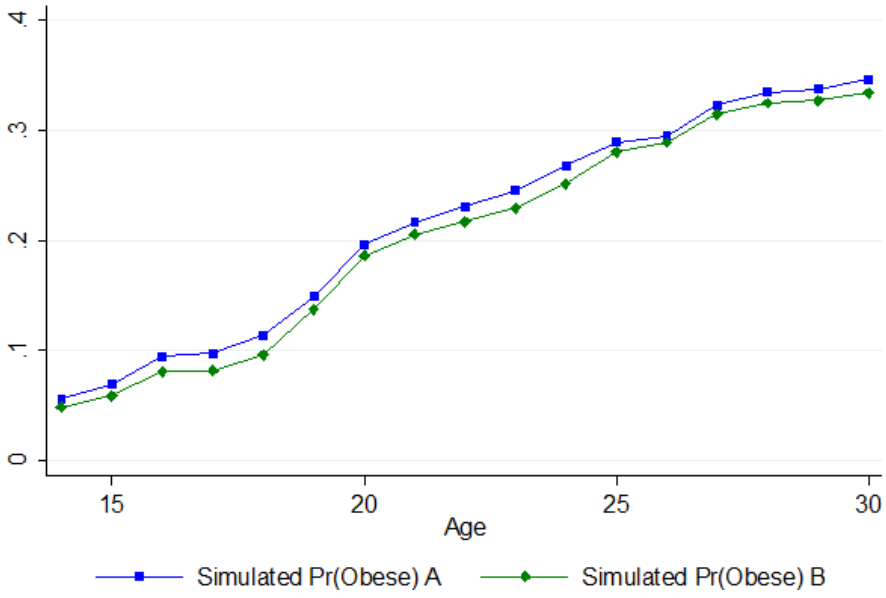


5.4.2 Simulated Marginal Changes of the Female Obesity Prevalence Rate

Using the estimated model and the simulation techniques described in this subsection, I simulate the changes in obesity prevalence as a result of changes in endogenous and exogenous variables. Three different situations are simulated using the estimated model. The first two relate to the individual decision to perform physical activity, and the last one has to do with the availability of physical- activity-related amenities in the individual's neighborhood. In the first exercise I simulate the female obesity prevalence rate that would have resulted from a state of the world in which all respondents perform high levels of physical activity while they are high school students. Such a situation could be the result of a policy that implements intense physical activity programs in schools nationwide. In the second exercise, an extension of the first one, I simulate the female obesity prevalence in a state of the world in which all respondents performed high levels of physical activity throughout their lives. This simulation describes an upper boundary of the role of physical activity as a partial solution for the critical obesity problem in the United States. In the last simulation I increase the availability of a set of neighborhood amenities that have an effect of increasing the probability of positive levels of physical activity. Then I simulate the female obesity prevalence that would have resulted in a state of the world in which these additional amenities are available to the respondents.

Intense Physical Activity in High School Figure 2 shows a comparison between the predicted obesity prevalence using the observed state of the world (A) and the predicted obesity prevalence in a state of the world where individuals perform high-level physical activity when they are in high school (B). The blue line represents the unaltered prediction of the female obesity prevalence rate and the green line represents the prediction when all females are assumed to perform intense PA during high school. As can be observed from the figure, the effect of intense PA in high school implies a reduction of the probability of obesity, and this reduction remains throughout the time the women are observed. This reduction is statistically significant.

Figure 3: Simulated Effects of Intense PA During High School



Simulated Effects in Wave IV

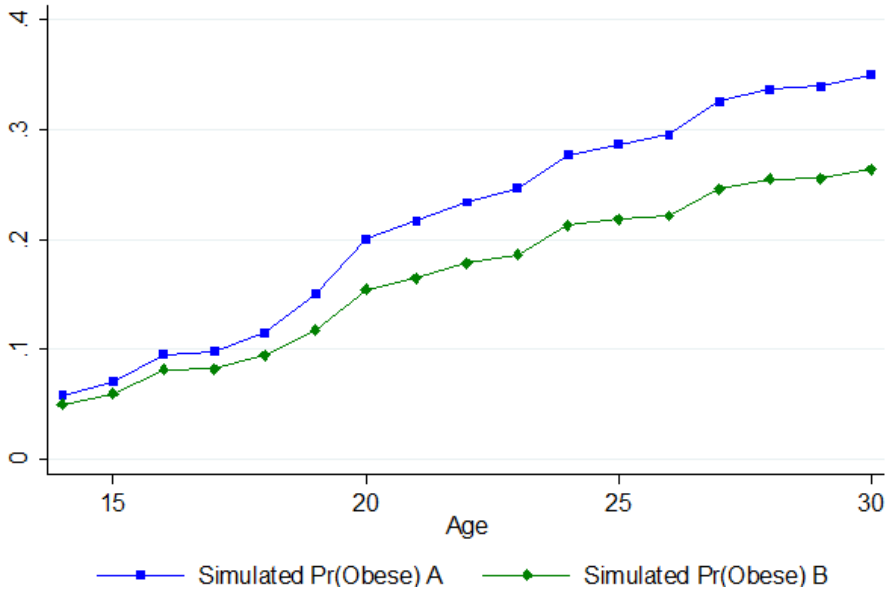
Simulations	Mean	Std. Dev.	T
1000	-0.01	0.003	-3.72

The small box presents a summary of these simulation results. This table presents the average effect of the simulations previously described for all individuals in the last wave of AddHealth. A generalized practice of intense PA during high school causes a reduction of 1.1% in the probability of being obese when these women are adults between 26 and 31 years old. In other words, given that previous weight status is a very important factor in explaining current weight status, good health practices such as intense PA during adolescence have a significant impact for women when they are in young adulthood.

Intense Physical Activity throughout Adolescence and Young Adulthood Figure 3 shows a comparison between the predicted obesity prevalence using the observed state of the world (A) and the

prediction in a state of the world in which individuals perform high-level physical activity throughout the years they are observed in the AddHealth study (B). The blue line represents the unaltered prediction of the female obesity prevalence rate, and the red line represents the prediction when all females in the sample are assumed to constantly perform intense PA during the entire period. As can be observed from the figure, the effect of intense PA in high school implies a great reduction of the probability of obesity. This reduction is statistically significant.

Figure 4: Simulated effects of constant intense PA



Simulated Effects in Wave IV

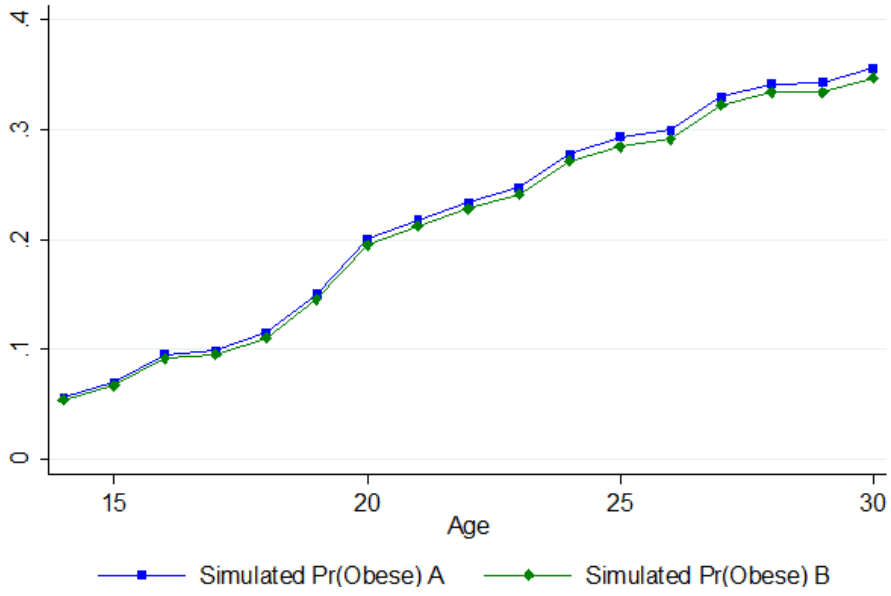
Simulations	Mean	Std. Dev.	T
1000	-0.08	0.02	-4.24

The small box presents the average effect of the simulation previously described for all female respondents in the last wave of AddHealth. A generalized practice of intense PA during the entire period that the women are observed causes a reduction of 9% in the probability of being obese when they are adults between 26 and 31 years old. This simulation is useful in terms of evaluating the potential that a strategy based on the encouragement of physical activity has on the reduction of obesity prevalence. The average effect of the simulation is a reduction of 8 percentage points in the prevalence of obesity. This is an important reduction, considering that the real obesity prevalence rate for women is close to 35%.

Increase in Physical Activity Related Neighborhood Amenities In this simulation-based exercise, I increase by one standard deviation a set of amenities that may encourage female AddHealth

respondents to perform positive levels of physical activity. The amenities in this set include the square mileage of public community parks, the number of parks in the neighborhood, the number of physical-activity-related facilities in the neighborhood that individuals can use by paying a fee, and the number of physical-activity-related facilities in the neighborhood for which some teaching and learning process is involved. Figure 4 shows a comparison between the predicted obesity prevalence before the increase in amenities (A) and the prediction after the increase in amenities (B).

Figure 5: Simulated Effects of One Standard Deviation Increase in Amenities



Simulated Effects in Wave IV

Simulations	Mean	Std. Dev.	T
1000	-0.01	0.004	-2.0

The box above presents the average effect of the simulation previously described for all female respondents in the last wave of AddHealth. An increase of one standard deviation in the PA-related amenities causes a reduction of almost 1 percentage point in the probability of being obese when these women are adults between 26 and 31 year old. This reduction is significant at 5% significance level.

5.5 Estimations Results of the Model for Men

5.5.1 Obesity Equation

In table 5 I present the results of the weight status estimation for the male sample. The first panel contains the results of an individual logit. In the other specifications, the obesity equation is jointly estimated together with all other equations in the system as well as initial conditions. In Specification

2, I use random sampling of the choice set of neighborhoods for the estimation of the residential location equation. In Specification 3, I use aggregation of the categories into type of neighborhoods for the estimation of the residential location equation. Both obesity equations in the second and third panels include three permanent and three time-varying unobserved heterogeneity parameters. Inclusion of additional unobserved heterogeneity parameters (permanent and time-varying) did not improve the performance of the models in terms of a significant reduction of the log likelihood function.

As in the case of females, lagged obesity is the main factor that explains the probability of obesity for males. Previous obesity increases of almost 70 percentage points to the probability of being obese in the present. In the individual logit, as well as in the jointly estimated specifications, physical activity is a significant factor in the reduction of the probability of being obese, but only for men who perform intense PA. In specification 3 the coefficient for intense PA is almost significant at 10% level, in the other cases the coefficient is strongly significant. In all these cases, there is a reduction of almost 3 percentage points in the probability of being obese. There is a small reduction in the coefficients of intense levels of PA in Specification 2 and 3 in comparison with the PA coefficients of the independent logit, especially in specification 3.

In comparison with the reference career-category, college-educated white collars workers, only two careers increase significantly the probability of obesity, low-educated blue collar workers and the low-educated white collars careers, however, in specification 3 the first one is not significant either. Smoking has a negative and significant effect in the determination of obesity for men that remains significant in the jointly estimated model. Obesity increases with age in a non-linear way. Other exogenous variables (e.g. the dummy for married, and the dummy for Hispanic ethnic background) significantly increase the probability of obesity. Being a first-generation immigrant and parental education greater than bachelor significantly reduce the probability of obesity in both the individual logit estimation and the jointly estimated model.

5.5.2 Input Equations Estimation Results for Men

In this subsection I present the estimation results of the three endogenous inputs that are contemporaneously included in the obesity equation for men: physical activity, smoking, and the number of fast-food meals per week. Table 6 shows the results of the physical activity equation estimation. The results for the other two inputs appear in table 7. As before, two specifications are presented in each table. Specification 1 is an independently estimated multinomial logit. In Specification 2, the PA equation is jointly estimated with all other equations in the system with random sampling of the choice set of neighborhoods.

Table 5: Weight Status Estimation Results for Men

Variable	[1]:Logit				[2]: Jointly estimated with Random Sampling of the Neighborhood Choice Set[1]				[3]: Jointly estimated with Agregation of the Neighborhood Choice Set[1]			
	Coef	S. D.		Mfx	Coef	S. D.		Mfx	Coef	S. D.		Mfx
Constant	-5.464	1.084	***		-5.422	1.051	***		-5.421	1.019	***	
Obese previous pd	3.833	0.092	***	0.696	3.824	0.092	***	0.695	3.832	0.093	***	0.696
Smoker	-0.149	0.065	**	-0.017	-0.155	0.065	***	-0.017	-0.159	0.067	***	-0.018
PA 1 or 2 times per week	-0.127	0.092		-0.014	-0.124	0.093		-0.014	-0.125	0.093		-0.014
PA 3 or 4 times per week	-0.124	0.084		-0.014	-0.120	0.086		-0.013	-0.111	0.086		-0.012
PA 5+ times per week	-0.271	0.082	***	-0.030	-0.260	0.083	***	-0.030	-0.220	0.147		-0.024
# Fast Food meals	0.011	0.013		0.001	0.007	0.025		0.001	0.019	0.013		0.002
Students	0.059	0.142		0.007	0.058	0.144		0.007	0.002	0.182		0.000
High Education/ Blue Collars	0.208	0.117	*	0.024	0.201	0.118	*	0.023	0.127	0.166		0.015
Med-Low Education/White Collars	0.469	0.127	***	0.058	0.467	0.128	***	0.058	0.351	0.183	*	0.042
Med-Low Education/Blue Collars	0.106	0.125		0.012	0.089	0.128		0.010	-0.055	0.201		-0.006
Med-Low Education/Not working	-0.008	0.153		-0.001	-0.020	0.157		-0.002	-0.168	0.225		-0.018
Age	0.233	0.097	***	0.030	0.235	0.090	***	0.031	0.234	0.091	***	0.031
Age^2	-0.005	0.002	***	-0.001	-0.005	0.002	***	-0.001	-0.005	0.002	***	-0.001
African American	0.045	0.086		0.005	0.046	0.086		0.005	0.037	0.086		0.004
Asian	-0.085	0.158		-0.009	-0.084	0.159		-0.009	-0.093	0.159		-0.010
Hispanic	0.338	0.096	***	0.040	0.344	0.097	***	0.041	0.339	0.096	***	0.041
1st generation immigrant	-0.353	0.171	**	-0.036	-0.357	0.173	**	-0.037	-0.354	0.173	**	-0.037
2nd generation immigrant	0.142	0.142		0.016	0.145	0.144		0.017	0.141	0.143		0.016
Married	0.367	0.100	***	0.044	0.367	0.101	***	0.044	0.376	0.101	***	0.045
Cohabitating	0.108	0.101		0.012	0.108	0.102		0.012	0.118	0.102		0.014
Divorced or separated	0.084	0.110		0.010	0.081	0.111		0.009	0.085	0.110		0.010
Living with Parents	-0.016	0.092		-0.002	-0.017	0.093		-0.002	-0.012	0.093		-0.001
No. Children younger than 6	0.049	0.060		0.006	0.050	0.061		0.006	0.052	0.061		0.006
No. Children older than 6	-0.064	0.093		-0.007	-0.065	0.094		-0.007	-0.063	0.094		-0.007
Family size	0.016	0.024		0.002	0.016	0.024		0.002	0.017	0.024		0.002
Initial H/H: Step parents	-0.047	0.092		-0.005	-0.046	0.092		-0.005	-0.037	0.093		-0.004
Initial H/H: Single Father	-0.330	0.195	*	-0.034	-0.325	0.197	*	-0.034	-0.308	0.197		-0.032
Initial H/H: Step Mother	-0.013	0.088		-0.002	-0.010	0.089		-0.001	-0.007	0.088		-0.001
Initial H/H: Non Parents	0.322	0.162	**	0.039	0.330	0.165	**	0.040	0.327	0.165	**	0.040
Parents Education: High School	0.217	0.117	*	0.025	0.217	0.117	*	0.025	0.203	0.118	*	0.023
Parents Education: Some College	0.010	0.117		0.001	0.012	0.118		0.001	-0.014	0.119		-0.002
Parents Education: Bachelor	0.052	0.126		0.006	0.054	0.127		0.006	0.014	0.130		0.002
Parents Education: +Bachelor	-0.280	0.139	**	-0.030	-0.282	0.140	**	-0.030	-0.335	0.146	**	-0.036
Parents Education: Missing	0.059	0.177		0.007	0.054	0.179		0.006	0.042	0.180		0.005
Dummy third wave	0.663	0.201	***	0.077	0.659	0.201	***	0.076	0.798	0.268	***	0.093
Dummy Fourth wave	0.905	0.262	***	0.111	0.897	0.266	***	0.110	1.014	0.340	***	0.126

Notes:

*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

[1]The definition of the neighborhood-type clusters is based on a K-medians cluster procedure with 5 categories

5.5.3 Physical Activity

The equation for physical activity (PA) is specified as a multinomial model with four categories that are associated with different intensity levels of physical activity (PA): no physical activity at all (1), one or two times per week (2), three to four times per week (3), and five or more times per week (4). The reference category is the first one (no physical activity at all).

Previous obesity reduces the probability of performing low, medium, and intense PA. This effect is significant in both of the specifications in table 6. There is a negative and significant effect of previous smoking on the probability of performing intense PA levels. Previous practice of PA increases the probability of PA in the present. For instance, individuals are more likely to perform intense PA if their past PA level has been at all positive and especially if it has been intense. In the jointly estimated model, the probability of intense PA increases by 12 percentage points for individuals who performed medium PA (3 to 4 times per week) during the previous period, and 31 percentage points for those who performed intense PA during the previous period. With the exception of students, almost all other careers significantly reduce the probability of positive PA levels in comparison with the reference category, college-educated white collar workers; however, highly-educated blue collar workers have a significant positive effect in the probability of intense PA in comparison with the reference category. There is a positive effect for African Americans in the probability of intense and madium PA and this effect remain significant in the jointly estimated model. Being a first-generation immigrant reduces the probability of intense PA by almost 6 percentage points. Similarly, being married or cohabitating reduce the probability of all PA levels; this effect is significant after controlling for unobserved heterogeneity in the jointly estimated model.

There is a significant reduction in the probability of intense PA with any additional children younger than six which is significant after controlling for unobserved heterogeneity. The probability of intense PA reduces with age; this reduction is significant even after controlling by endogeneity. In summary, male respondents in AddHealth are significantly more likely to perform intense PA during the current period if they did not smoke and if they practiced medium or intense PA during the previous period. As before, this evidence suggests that healthy behaviors in the past increase the probability of healthy behaviors in the present. After controlling for endogeneity two neighborhood-amenity variables has a positive and significant effect in the probability of intense PA levels. The number of public-PA-related facilities and the number of Fee required PA-related facilities in the neighborhood have a positive and significant effect in the probability of intense PA.

Table 6: Physical Activity Estimation Results for Men

Variable	[1]: Logit PA=2 relative to PA=1			[2]: Jointly Estimated PA=2 relative to PA=1			[1]: Logit PA=3 relative to PA=1			[2]: Jointly Estimated PA=3 relative to PA=1			[1]: Logit PA=4 relative to PA=1			[2]: Jointly Estimated PA=4 relative to PA=1		
	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx
Constant	1.153	1.645		1.128	1.481		1.673	1.448		3.534	1.636 **		7.233	1.405 ***		7.685	1.584 ***	
Obese (t-1)	-0.243	0.097 ***	-0.014	-0.244	0.099 ***	-0.014	-0.189	0.089 **	-0.007	-0.181	0.102 *	-0.005	-0.200	0.088 **	-0.011	-0.206	0.105 *	-0.013
Smoker (t-1)	0.030	0.071	0.017	0.032	0.072	0.017	-0.063	0.066	0.013	-0.067	0.074	0.012	-0.282	0.065 ***	-0.047	-0.281	0.077 ***	-0.047
PA 1 or 2 times per week (t-1)	0.716	0.112 ***	0.029	0.716	0.115 ***	0.028	0.702	0.113 ***	0.028	0.710	0.127 ***	0.029	0.750	0.125 ***	0.043	0.753	0.141 ***	0.043
PA 3 or 4 times per week (t-1)	0.710	0.104 ***	-0.009	0.711	0.106 ***	-0.010	1.091	0.099 ***	0.037	1.094	0.112 ***	0.037	1.370	0.108 ***	0.122	1.370	0.121 ***	0.122
PA 5+ times per week (t-1)	0.854	0.104 ***	-0.052	0.858	0.106 ***	-0.054	1.462	0.099 ***	-0.001	1.475	0.109 ***	0.000	2.495	0.104 ***	0.313	2.503	0.115 ***	0.315
# Fast Food meals (t-1)	0.003	0.017	0.000	0.002	0.017	0.000	0.005	0.015	0.001	0.006	0.017	0.001	-0.005	0.015	-0.001	0.002	0.017	0.000
Students	-0.083	0.153	-0.011	-0.087	0.155	-0.011	-0.092	0.141	-0.024	-0.109	0.157	-0.026	0.125	0.138	0.034	0.120	0.153	0.035
High Education/ Blue Collars	-0.486	0.127 ***	-0.024	-0.481	0.128 ***	-0.024	-0.578	0.118 ***	-0.050	-0.570	0.128 ***	-0.051	-0.356	0.117 ***	0.002	-0.338	0.129 ***	0.005
Med-Low Education/White Collars	-0.563	0.142 ***	-0.024	-0.552	0.145 ***	-0.025	-0.632	0.132 ***	-0.042	-0.616	0.146 ***	-0.043	-0.542	0.132 ***	-0.025	-0.502	0.146 ***	-0.020
Med-Low Education/Blue Collars	-0.676	0.134 ***	-0.020	-0.667	0.138 ***	-0.020	-0.846	0.126 ***	-0.049	-0.828	0.141 ***	-0.049	-0.866	0.127 ***	-0.061	-0.833	0.141 ***	-0.057
Med-Low Education/Not working	-0.762	0.163 ***	-0.029	-0.770	0.167 ***	-0.029	-1.008	0.156 ***	-0.076	-1.025	0.172 ***	-0.077	-0.772	0.151 ***	-0.032	-0.782	0.169 ***	-0.033
Age	0.008	0.120	0.029	-0.001	0.114	0.028	-0.205	0.105 *	-0.005	-0.204	0.118 *	-0.005	-0.593	0.102 ***	-0.061	-0.601	0.115 ***	-0.061
Age^2	0.000	0.002	0.000	0.000	0.002	0.000	0.005	0.002 **	0.000	0.004	0.002 *	0.000	0.012	0.002 ***	0.001	0.012	0.002 ***	0.001
African American	0.074	0.101	-0.008	0.078	0.102	-0.008	0.287	0.088 ***	0.031	0.287	0.103 ***	0.030	0.190	0.087 **	0.006	0.194	0.104 *	0.007
Asian	0.011	0.179	-0.014	0.012	0.186	-0.015	0.297	0.160 *	0.039	0.327	0.181 *	0.044	0.146	0.158	-0.001	0.156	0.179	-0.003
Hispanic	0.093	0.116	-0.002	0.093	0.119	-0.002	0.214	0.105 **	0.022	0.227	0.121 *	0.024	0.138	0.103	0.002	0.140	0.120	0.001
1st generation immigrant	-0.009	0.182	0.013	-0.005	0.187	0.013	-0.078	0.167	0.012	-0.060	0.192	0.015	-0.291	0.168 *	-0.045	-0.288	0.191	-0.047
2nd generation immigrant	0.131	0.166	0.006	0.132	0.170	0.006	0.161	0.152	0.015	0.159	0.175	0.015	0.102	0.149	-0.001	0.094	0.171	-0.002
Married	-0.262	0.111 ***	0.002	-0.269	0.113 ***	0.002	-0.374	0.106 ***	-0.010	-0.376	0.120 ***	-0.009	-0.537	0.106 ***	-0.055	-0.551	0.124 ***	-0.057
Cohabitating	-0.364	0.108 ***	-0.004	-0.370	0.110 ***	-0.005	-0.479	0.102 ***	-0.017	-0.465	0.115 ***	-0.014	-0.624	0.102 ***	-0.058	-0.635	0.120 ***	-0.061
Divorced or separated	-0.097	0.119	-0.013	-0.095	0.121	-0.013	-0.098	0.112	-0.025	-0.099	0.126	-0.025	0.126	0.105	0.036	0.133	0.126	0.037
Living with Parents	-0.013	0.099	0.007	-0.014	0.101	0.007	-0.145	0.092	-0.014	-0.147	0.107	-0.014	-0.114	0.090	-0.008	-0.116	0.108	-0.008
No. Children younger than 6	0.023	0.066	0.015	0.018	0.066	0.015	-0.102	0.066	0.001	-0.108	0.073	0.001	-0.223	0.070 ***	-0.032	-0.229	0.081 ***	-0.033
No. Children older than 6	0.080	0.100	0.003	0.083	0.098	0.004	0.104	0.097	0.010	0.100	0.109	0.009	0.066	0.105	-0.001	0.067	0.123	0.000
Family size	0.020	0.026	0.003	0.020	0.026	0.003	0.007	0.024	0.002	0.004	0.027	0.001	-0.014	0.023	-0.004	-0.015	0.028	-0.004

Continuing in the next page

Table 6: Physical Activity Estimation Results for Men (Continued from previous page)

Variable	[1]: Logit PA=2 relative to PA=1			[2]: Jointly Estimated PA=2 relative to PA=1			[1]: Logit PA=3 relative to PA=1			[2]: Jointly Estimated PA=3 relative to PA=1			[1]: Logit PA=4 relative to PA=1			[2]: Jointly Estimated PA=4 relative to PA=1		
	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx
Initial H/H: Step parents	-0.002	0.100	-0.003	-0.001	0.101	-0.003	0.119	0.091	0.023	0.112	0.105	0.022	-0.025	0.090	-0.015	-0.023	0.108	-0.014
Initial H/H: Single Father	-0.218	0.205	-0.024	-0.223	0.208	-0.024	0.113	0.172	0.033	0.107	0.192	0.033	-0.068	0.174	-0.015	-0.072	0.190	-0.015
Initial H/H: Step Mother	-0.018	0.098	-0.009	-0.014	0.099	-0.009	0.156	0.087 *	0.023	0.167	0.102 *	0.024	0.049	0.086	-0.005	0.061	0.102	-0.004
Initial H/H: Non Parents	0.111	0.197	-0.008	0.113	0.204	-0.008	0.333	0.173 *	0.032	0.330	0.203	0.031	0.250	0.171	0.011	0.257	0.197	0.013
Parents Education: High School	-0.172	0.126	-0.027	-0.174	0.132	-0.027	0.124	0.120	0.015	0.123	0.133	0.015	0.131	0.118	0.019	0.123	0.132	0.018
Parents Education: Some College	-0.164	0.126	-0.026	-0.166	0.131	-0.026	0.114	0.120	0.015	0.110	0.132	0.015	0.108	0.118	0.016	0.103	0.131	0.015
Parents Education: Bachelor	-0.187	0.136	-0.028	-0.193	0.141	-0.028	0.184	0.128	0.033	0.187	0.141	0.035	0.055	0.126	0.000	0.047	0.140	-0.002
Parents Education: +Bachelor	-0.016	0.146	-0.016	-0.022	0.153	-0.017	0.278	0.137 **	0.036	0.285	0.154 *	0.039	0.146	0.135	0.002	0.137	0.153	0.000
Parents Education: Missing	-0.240	0.200	-0.031	-0.239	0.208	-0.032	0.023	0.182	-0.002	0.044	0.207	0.001	0.139	0.177	0.032	0.146	0.202	0.032
Ground Transportation Terminals by County/ Area	0.003	0.224	0.013	0.016	0.229	0.011	-0.153	0.201	-0.009	-0.097	0.229	-0.004	-0.202	0.190	-0.023	-0.157	0.215	-0.021
Square miles of parks within 1km of Tract boundaries	-0.006	0.027	-0.002	-0.007	0.027	-0.002	0.025	0.023	0.003	0.025	0.023	0.003	0.019	0.024	0.002	0.018	0.025	0.001
Beta Street conectivity index within 5km Buffers	-0.577	0.364	-0.054	-0.583	0.382	-0.056	0.033	0.329	0.021	0.034	0.370	0.019	-0.003	0.318	0.014	0.018	0.361	0.019
Parks within 3km Buffers	0.000	0.073	0.006	0.001	0.075	0.007	-0.034	0.065	0.006	-0.045	0.074	0.004	-0.136	0.064 **	-0.022	-0.131	0.077 *	-0.020
Public PA related amenities within 5km buffers	0.082	0.135	-0.002	0.089	0.141	-0.001	0.069	0.121	-0.010	0.066	0.140	-0.012	0.218	0.117 *	0.031	0.231	0.138 *	0.034
Fee required PA related Amenities 5km buffers	0.138	0.142	-0.001	0.142	0.145	-0.002	0.171	0.128	-0.001	0.178	0.148	-0.002	0.288	0.124 **	0.033	0.303	0.147 **	0.035
Non PA related Amenities 5km buffers	-0.121	0.117	-0.003	-0.121	0.118	-0.002	-0.133	0.103	-0.003	-0.133	0.120	-0.002	-0.182	0.097 *	-0.017	-0.190	0.113 *	-0.018
Instruction PA related amenities 5km buffers	0.058	0.064	0.002	0.059	0.065	0.003	0.044	0.058	-0.001	0.042	0.066	-0.001	0.071	0.055	0.007	0.070	0.067	0.007
Membership required PA amenities 5km buffers	-0.135	0.099	-0.008	-0.146	0.101	-0.008	-0.083	0.088	0.000	-0.099	0.102	-0.001	-0.112	0.086	-0.008	-0.135	0.102	-0.010
Outdoor PA related amenities 5km buffers	0.062	0.147	-0.002	0.066	0.154	-0.003	0.167	0.132	0.018	0.165	0.154	0.016	0.103	0.129	0.001	0.123	0.152	0.005
Amusement Park PA related amenities 5km buffers	-0.287	0.545	0.102	-0.310	0.585	0.102	-1.830	0.517 ***	-0.119	-1.864	0.567 ***	-0.120	-2.101	0.498 ***	-0.198	-2.163	0.551 ***	-0.204
ACCRA price of a cigarrete Carton, 2005 dollars	-0.009	0.011	-0.001	-0.009	0.011	-0.001	-0.002	0.010	0.000	-0.002	0.011	0.000	0.000	0.009	0.000	-0.001	0.011	0.000
ACCRA Index price for Groceries, 2005 dollars	-0.490	0.263 *	-0.044	-0.495	0.279 *	-0.043	0.147	0.242	0.065	0.132	0.267	0.067	-0.247	0.235	-0.043	-0.291	0.261	-0.050
ACCRA Index price for Junk food, 2005 dollars	-0.089	0.146	-0.008	-0.085	0.149	-0.008	0.055	0.129	0.021	0.047	0.153	0.019	-0.092	0.127	-0.018	-0.090	0.151	-0.018
ACCRA cost of living Index price, 2005 dollars	0.897	0.340 ***	0.066	0.895	0.371 ***	0.064	0.277	0.307	-0.060	0.260	0.348	-0.064	0.845	0.299 ***	0.090	0.864	0.339 ***	0.096
Dummy third wave	-0.478	0.321	0.061	-0.432	0.341	0.060	-1.696	0.291 ***	-0.129	-1.600	0.330 ***	-0.124	-1.571	0.284 ***	-0.121	-1.457	0.319 ***	-0.111
Dummy Fourth wave	-0.578	0.369	0.014	-0.527	0.397	0.016	-1.503	0.336 ***	-0.154	-1.459	0.382 ***	-0.155	-0.881	0.329 ***	-0.016	-0.796	0.374 **	-0.007

*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

5.5.4 Smoking, Childbirth, and Fast Food Meals for Men

Table 7 presents the results of the estimation of the equations for the smoking decision and the linear model for the number of fast-food meals. As before, two specifications are presented for each equation. Specification 1 is the estimation of each equation separately, and Specification 2 represents each equation estimated jointly with all remaining equations of the system. Male respondents in AddHealth are more likely to smoke regularly during the current period if they smoked during the previous period; this probability also increase with the number of fast food meals they reported. After controlling for unobserved heterogeneity, all of these relationships remain significant. As with previous results, this seems to suggest that unhealthy practices in the past explain current unhealthy lifestyles. For previous smokers there is an increment of 47 percentage points in the probability of smoking. Compared to college-educated white collar workers, except by students, all other careers have a significant positive impact in the probability of smoking for men.

Some demographic variables have a negative effect upon the probability of smoking, even after controlling for unobserved heterogeneity. The dummy variables for African American, Hispanic, and first-generation immigrant are negative, which implies a reduction in the probability of smoking in comparison with the reference categories (white males and third generation immigrant, repectively). There is a negative effect for married male respondents upon the probability of smoking in comparison with single females; this effect is also significant in the jointly estimated model. For separate or divorced females the effect is positive, however, which significantly increases their probability of smoking. Male respondents are more likely to smoke if they were observed at the beginning of the study in households with at least one step-parent or a single father, compared to men who were observed at the beginning of the study with two biological parents.

Male AddHealth responders consume more fast-food meals per week if they smoke, however, this effect is not significant after controlling for unobserved heterogeneity. Previous consumption of fast-food meals greatly explains current consumption. On average, African American males consume more fast-food meals than white males do. Male respondents who still live with their parents significantly consume more fast-food meals per week, whereas married and cohabitating males consume fewer fast-food meals per week on average. In addition, men whose parents are college-educated consume fewer fast-food meals per week.

Table 7: Smoking and Fast Food Meals Estimation Results

Variable	Smoking: Logit [1]			Smoking: Jointly Estimated [2]			Fast Food Meals: OLS [1]		Fast Food Meals: Jointly Estimated [2]		
	Coef	S. D.	Mfx	Coef	S. D.	Mfx	Coef	S. D.	Coef	S. D.	
Constant	-0.589	0.943		-0.134	0.993		0.055	0.775	10.697	0.631	***
Obese (t-1)	-0.008	0.073	-0.001	-0.013	0.074	-0.002	-0.090	0.060	-0.011	0.057	
Smoker (t-1)	2.206	0.050 ***	0.473	2.202	0.051 ***	0.473	0.117	0.044 ***	0.038	0.043	
PA 1 or 2 times per week (t-1)	-0.146	0.095	-0.024	-0.145	0.097	-0.024	0.058	0.079	0.080	0.071	
PA 3 or 4 times per week (t-1)	-0.092	0.083	-0.015	-0.094	0.084	-0.015	-0.031	0.068	0.004	0.062	
PA 5+ times per week (t-1)	-0.125	0.080	-0.021	-0.128	0.081	-0.021	0.072	0.066	0.109	0.059	*
# Fast Food meals (t-1)	0.031	0.012 ***	0.005	0.027	0.013 **	0.005	0.470	0.010 ***	0.408	0.013	***
Students	0.305	0.130 ***	0.052	0.311	0.132 ***	0.053	-0.047	0.098	-0.051	0.087	
High Education/ Blue Collars	0.749	0.108 ***	0.130	0.741	0.110 ***	0.129	0.327	0.083 ***	0.181	0.076	***
Med-Low Education/White Collars	0.398	0.122 ***	0.068	0.378	0.125 ***	0.065	0.552	0.093 ***	0.166	0.087	*
Med-Low Education/Blue Collars	1.055	0.114 ***	0.189	1.030	0.116 ***	0.185	0.638	0.088 ***	0.314	0.081	***
Med-Low Education/Not working	1.141	0.134 ***	0.206	1.136	0.138 ***	0.206	0.178	0.107 *	0.151	0.098	
Age	0.028	0.066	0.005	0.029	0.067	0.005	0.171	0.054 ***	0.209	0.044	***
Age2	-0.001	0.001	0.000	-0.001	0.002	0.000	-0.004	0.001 ***	-0.004	0.001	***
African American	-0.380	0.072 ***	-0.062	-0.396	0.073 ***	-0.064	0.362	0.058 ***	0.241	0.053	***
Asian	-0.011	0.129	-0.002	-0.028	0.132	-0.005	0.125	0.102	0.116	0.089	
Hispanic	-0.151	0.083 *	-0.025	-0.163	0.085 *	-0.027	0.115	0.068 *	0.115	0.063	*
1st generation immigrant	-0.320	0.145 **	-0.051	-0.340	0.148 **	-0.055	0.002	0.110	-0.042	0.096	
2nd generation immigrant	-0.135	0.122	-0.022	-0.147	0.125	-0.024	0.119	0.097	0.081	0.084	
Married	-0.349	0.093 ***	-0.056	-0.333	0.094 ***	-0.054	-0.342	0.074 ***	-0.175	0.067	***
Cohabitating	0.130	0.086	0.022	0.144	0.087 *	0.024	-0.348	0.072 ***	-0.124	0.067	*
Divorced or separated	0.482	0.093 ***	0.083	0.478	0.094 ***	0.083	-0.071	0.078	-0.124	0.073	*
Living with Parents	0.044	0.077	0.007	0.051	0.078	0.008	0.081	0.063	0.131	0.059	**
No. Children younger than 6	-0.039	0.056	-0.006	-0.032	0.056	-0.005	-0.018	0.045	0.059	0.043	
No. Children older than 6	0.117	0.082	0.020	0.113	0.083	0.019	0.162	0.068 ***	0.112	0.064	*
Family size	0.054	0.019 ***	0.009	0.053	0.019 ***	0.009	0.005	0.015	-0.007	0.015	

Continuing in the next page

Table 7: Smoking, and Fast Food Meals (Continued from previous page)

Variable	Smoking: Logit [1]			Smoking: Jointly Estimated [2]			Fast Food Meals: OLS [1]			Fast Food Meals: Jointly Estimated [2]		
	Coef	S. D.	Mfx	Coef	S. D.	Mfx	Coef	S. D.		Coef	S. D.	
Initial H/H: Step parents	0.305	0.071 ***	0.052	0.298	0.072 ***	0.051	0.130	0.060 **		0.035	0.054	
Initial H/H: Single Father	0.236	0.139 *	0.040	0.239	0.141 *	0.041	-0.132	0.117		-0.118	0.102	
Initial H/H: Step Mother	0.075	0.069	0.013	0.075	0.070	0.012	0.106	0.057 *		0.059	0.055	
Initial H/H: Non Parents	0.156	0.138	0.026	0.148	0.141	0.025	0.250	0.114 **		0.090	0.100	
Parents Education: High School	0.076	0.095	0.013	0.077	0.096	0.013	-0.136	0.078 *		-0.072	0.070	
Parents Education: Some College	0.092	0.095	0.015	0.091	0.096	0.015	-0.093	0.078		-0.058	0.069	
Parents Education: Bachelor	0.158	0.101	0.026	0.161	0.103	0.027	-0.185	0.083 **		-0.099	0.073	
Parents Education: +Bachelor	0.010	0.109	0.002	0.005	0.112	0.001	-0.197	0.088 **		-0.136	0.078 *	
Parents Education: Missing	0.050	0.144	0.008	0.051	0.147	0.008	-0.321	0.118 ***		-0.258	0.103 ***	
Ground Transportation Terminals by County/ Area	0.124	0.161	0.021	0.102	0.163	0.017	0.173	0.130		-0.002	0.111	
Square miles of parks within 1km of Tract boundaries	-0.039	0.023 *	-0.007	-0.037	0.023	-0.006	-0.022	0.016		0.003	0.016	
Beta Street conectivity index within 5km Buffers	-0.324	0.251	-0.052	-0.324	0.256	-0.052	0.351	0.205 *		0.063	0.172	
Parks within 3km Buffers	-0.149	0.054 ***	-0.024	-0.134	0.056 ***	-0.022	-0.040	0.042		-0.046	0.040	
Public PA related amenities within 5km buffers	0.106	0.099	0.018	0.109	0.101	0.018	0.211	0.078 ***		0.120	0.070 *	
Fee required PA related Amenities 5km buffers	-0.072	0.105	-0.012	-0.069	0.106	-0.011	0.017	0.084		-0.031	0.074	
Non PA related Amenities 5km buffers	-0.113	0.084	-0.019	-0.128	0.086	-0.021	0.014	0.068		0.000	0.062	
Instruction PA related amenities 5km buffers	0.083	0.047 *	0.014	0.086	0.048 *	0.014	-0.012	0.038		0.015	0.035	
Membership required PA amenities 5km buffers	0.039	0.072	0.006	0.045	0.073	0.008	-0.176	0.058 ***		-0.066	0.052	
Outdoor PA related amenities 5km buffers	0.001	0.109	0.000	0.029	0.113	0.005	0.076	0.087		0.021	0.077	
Amusement Park PA related amenities 5km buffers	0.349	0.428	0.060	0.323	0.440	0.055	-0.820	0.338 ***		-0.512	0.282 *	
ACCRA price of a cigarrete Carton, 2005 dollars	-0.005	0.007	-0.001	-0.005	0.008	-0.001	-0.013	0.006 **		-0.008	0.005	
ACCRA Index price for Groceries, 2005 dollars	0.297	0.193	0.050	0.285	0.198	0.048	-0.595	0.158 ***		-0.351	0.138 ***	
ACCRA Index price for Junk food, 2005 dollars	-0.086	0.096	-0.014	-0.079	0.099	-0.013	0.006	0.079		0.024	0.069	
ACCRA cost of living Index price, 2005 dollars	-0.555	0.206 ***	-0.087	-0.539	0.212 ***	-0.085	0.222	0.166		0.062	0.139	
Dummy third wave	-0.722	0.219 ***	-0.115	-0.773	0.224 ***	-0.123	0.418	0.176 ***		-0.281	0.152 *	
Dummy Fourth wave	-0.178	0.254	-0.029	-0.243	0.259	-0.040	0.100	0.205		-0.624	0.178 ***	

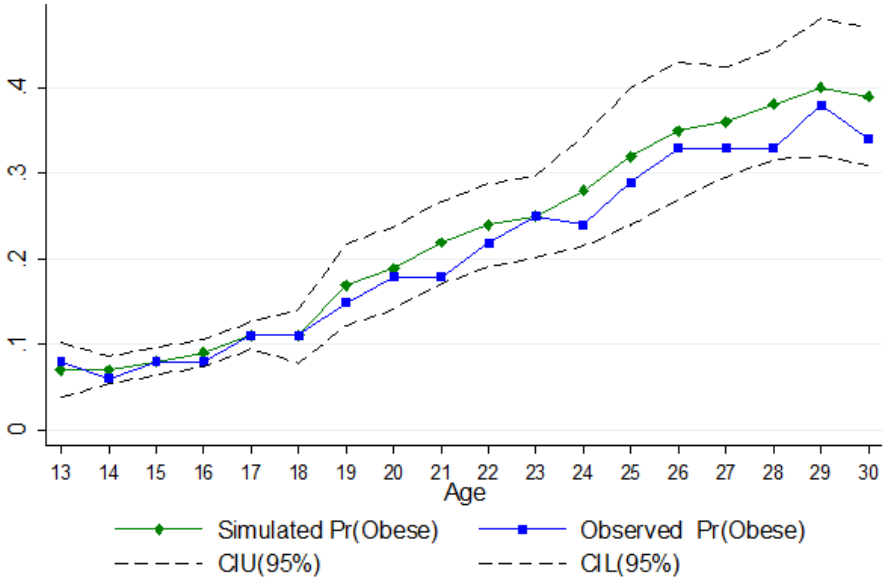
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

5.6 Simulations Using Parametric Bootstrap and Fit of the Model for Men

5.6.1 Fit of the Model for Males

Using the procedure described in the introduction of subsection 8.3, I predict the male obesity prevalence rate. Figure 6 presents the predictions and confidence intervals (at a 95% significance level) of the obesity prevalence rate at all ages at which respondents are observed in the AddHealth study. The green line represents the obesity prevalence rate estimated, averaged through all bootstrap samples, for the jointly estimated model (Specification 2). The blue line represents the obesity prevalence rate computed using the all the respondents in AddHealth observed at any of the specific ages represented in the horizontal axis. The upper and lower limits of the confidence intervals are represented by the black dotted lines. On average, the model predictions of the male obesity prevalence rate are a bit higher than the observed prevalence rate for male AddHealth respondents. In general the model captures well the evolution of obesity prevalence.

Figure 6: Model Predictions of Male Obesity Prevalence



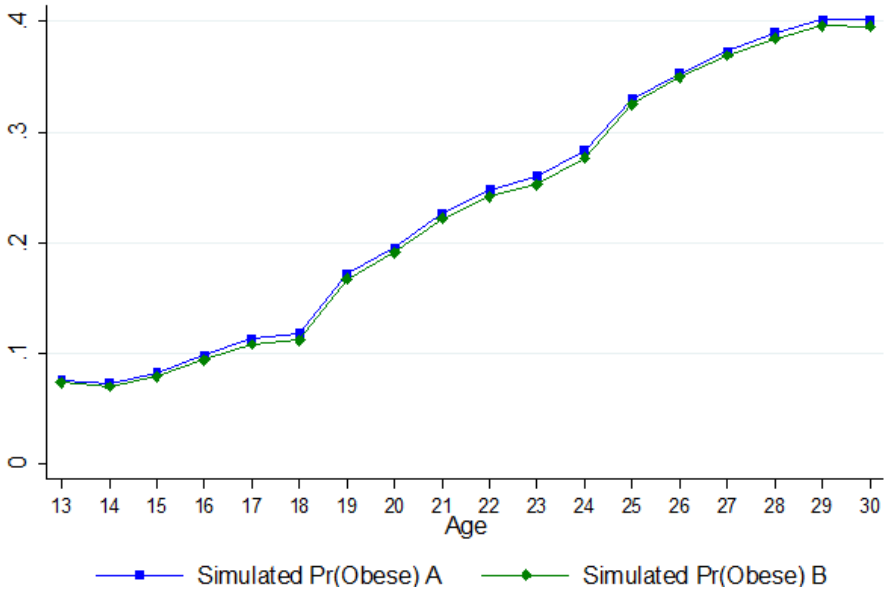
5.6.2 Simulated Marginal Changes in the Male Obesity Prevalence Rate

In this section, I present changes in the male obesity prevalence that result from three simulation exercises (the same exercises that were presented for women). In the first exercise I simulate the male obesity prevalence rate that would have resulted from a state of the world in which all respondents perform high levels of physical activity when they are high school students. In the second exercise, which is an extension of the first one, I simulate the male obesity prevalence in a state of the world in which all respondents perform high levels of physical activity throughout their lives. In the last simulation I increase the availability of a set of neighborhood amenities that have an effect upon increasing the probability of positive levels of physical activity. Then I simulate the male obesity prevalence that would have resulted in a state of the world in which these additional amenities are available for the male respondents.

Intense Physical Activity in High School Figure 7 shows a comparison between the predicted obesity prevalence using the observed state of the world (A) and the prediction in a state of the world where individuals perform high-level physical activity when they are in high school (B). The blue line represents the unaltered prediction of the male obesity prevalence rate, and the green line represents the prediction when all men are assumed to perform intense PA during high school. As can be observed from the figure, the effect of intense PA in high school implies a very small reduction of the probability of obesity for men.

The small box below contains a summary of these simulation results. This table presents the average effect of the simulation previously described for all men in the last waive of Addhealth. A generalized practice of intense PA during high school causes a reduction of 0.5 percentage points in the probability of being obese when these men are adults between 26 and 31 years old. This reduction is statistically significant.

Figure 7: Simulated Effects of Intense PA During High School

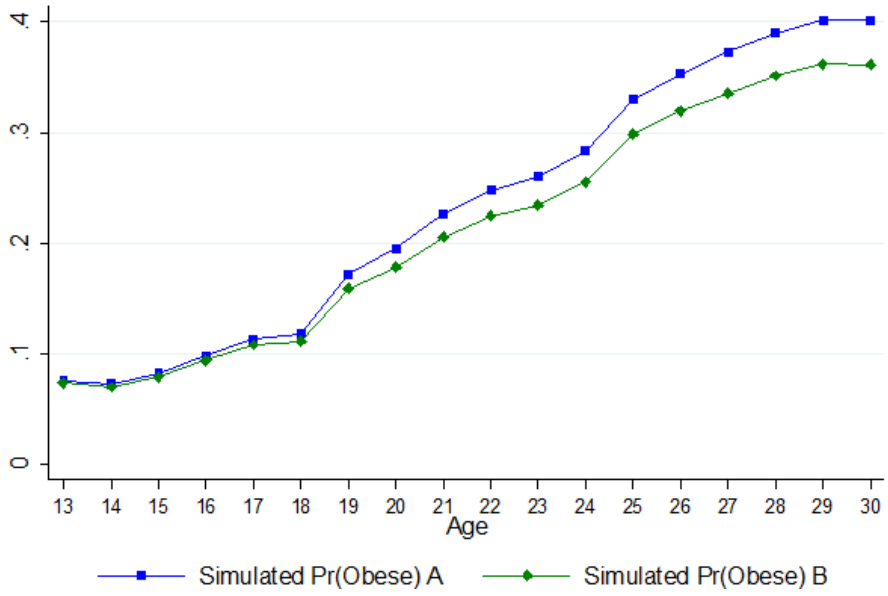


Simulated Effects in Wave IV

Simulations	Mean	Std. Dev.	T
1000	-0.005	0.002	-2.23

Intense Physical Activity Throughout Adolescence and Young Adulthood Figure 8 shows a comparison between the predicted obesity prevalence using the observed state of the world (A) and the prediction of a state of the world in which men perform high-level physical activity throughout all years they are observed in the AddHealth study (B). The blue line represents the unaltered prediction of the male obesity prevalence rate, and the green line represents the prediction when all males in the sample are assumed to perform intense PA constantly during the entire period. As can be seen in the figure, the effect of constant intense PA implies an important reduction of the probability of obesity. The small box presents the average effect of the simulation previously described for all male respondents in the last wave of AddHealth. A generalized practice of intense PA during the entire period that men are observed causes a reduction of 3.8 percentage points in the probability of being obese when these men are adults between 26 and 31 years old. This reduction is strongly statistically significant.

Figure 8: Simulation Effects of Constant Intense PA



Simulated Effects in Wave IV

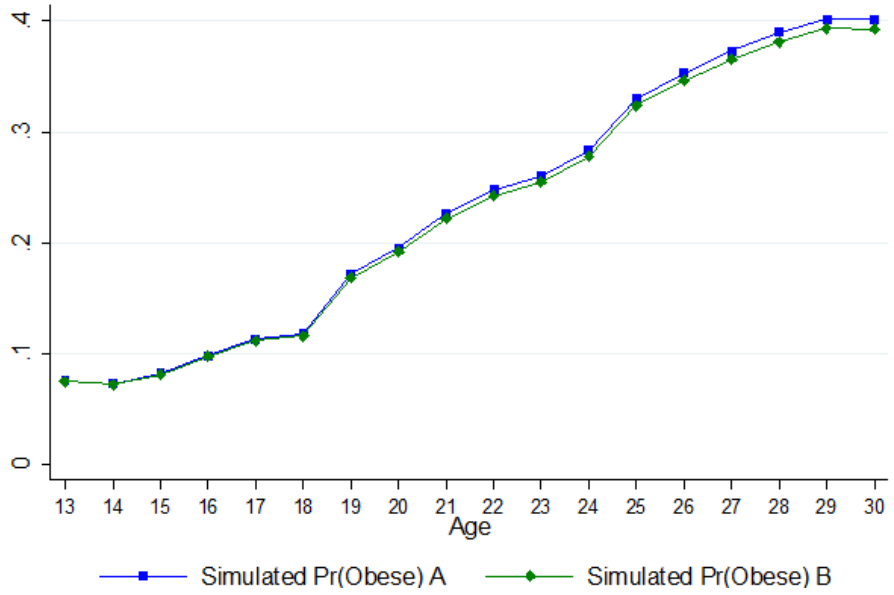
Simulations	Mean	Std. Dev.	T
1000	-0.038	0.015	-2.46

Increase in Physical Activity Related Neighborhood Amenities In this simulation-based exercise, I increase by one standard deviation a set of amenities that may encourage the male AddHealth respondent to perform positive levels of physical activity. The amenities in this set are: the square mileage of public community parks, the number of physical-activity-related facilities in the neighborhood that individuals can use by paying a fee, the number of public physical-activity-related facilities in the neighborhood, the number of physical-activity-related facilities in the neighborhood for which some teaching and learning process is involved, and the number of outdoor physical-activity-related facilities in the neighborhood. Figure 9 shows a comparison between the predicted obesity prevalence before the increment in amenities (A) and the prediction after the increment in amenities (B). The small box below contains the average effect of the simulation previously described for all male respondents in the last wave of AddHealth. An increase of one standard deviation in the PA-related amenities causes a significant reduction of 1 percentage point in the probability of being obese when these men are adults between 26 and 31 years old.

5.7 Remarks on the Validity of Exclusion Restrictions

In the estimated equations presented in the previous sections several variables were excluded from the structural equation for weight status. These variables were included in other equations of the system,

Figure 9: Simulatied Effects of One Standard Deviation Increase in Amenities



Simulated Effectsin Wave IV

Simulations	Mean	Std. Dev.	T
1000	-0.01	0.004	-2.11

most of them are neighborhood characteristics and local prices (i.e., variables in vector I_{it} in equations 5.3 to 5.6). These variables play the role of exclusions restrictions that contribute to the identification of the system. They are supposed to have only an indirect impact on obesity through the effect they have on all other endogenous choices. One way to test this assumption is using log-likelihood ratio test to test the hypothesis of jointly significance of these variables in the reduced form equations for endogenous choices and in the structural equation of weight determination. The null hypothesis of a log-likelihood ratio test is that a subset of coefficients in the estimation is jointly equal to zero or $H_o : \beta_1^E = \beta_2^E = ... = \beta_h^E$, where β^E is the coefficient for a generic exclusion restriction and there are h exclusion restrictions in total. Under the null hypothesis the test statistics is $2(L_u - L_r)$, which is distributed χ^2 with h degrees of freedom, with L_u and L_r the loglikelihood values of the unrestricted and restricted model respectively.

The null hypothesis of jointly insignificance of the exclusions restrictions in all the endogenous choice equations is strongly rejected for the models of women and men. In the case of women the value of the log-likelihood function of the model presented in subsection 6.1.2 was -90267, and the value of the log-likelihood function for a restricted model where all coefficients of exclusions restriction are equal to zero was -90388. These values for the men's estimation were -78535 and -76317 respectively. With these values, and their respective degrees of freedom[14] the null hypothesis of the test is rejected at very low p-values. From this I conclude that all exclusion restrictions are important explanatory factors of the

[14] There are a total number of 90 exclusion restrictions for women and 75 for men in the system.

endogenous decisions. In addition, I perform similar tests using the estimated models presented in the previous subsections as restricted models and as unrestricted models specifications in which I included all exclusion restrictions in the structural equation for obesity. For men and women I obtain that the likelihood function does not reduces in the unrestricted model, therefore the null hypothesis cannot be rejected at any reasonable level. I conclude from this that exclusion restrictions are not important explanatory factors in the structural equation.

6 Conclusion and Final Remarks

6.1 Conclusion

In this research I propose a comprehensive model of weight status determination. The model is estimated separately for a sample of female and male respondents in the AddHealth study. In this model weight status is modeled as a health outcome derived from a series of decisions that are treated as endogenous. The endogenous decisions modeled in this research are lifestyle decisions and major individual decisions. Lifestyle decisions represent the behavior of the individuals with regard to practices that could directly or indirectly modify the biological process through which weight status is determined. Among these lifestyle decisions, this research places special emphasis upon physical activity, which is a measure of the intensity of an individual's energy expenditure during leisure time. Other lifestyle decisions modeled in this dissertation are: the smoking decision, a proxy for food consumption, and (for women) childbearing.

With regard to the major individual decisions, in this research I specifiy econometric models for residential location decisions, in terms of the neighborhood in which an individual decides to live, and career-related decisions, in terms of a set of categories that summarize individual main activity or occupation. I incorporate the residential location decision into the system as a way for controlling for the endogenous nature of the neighborhood characteristics. These characteristics capture neighborhood amenities and facilities that presumably may encourage healthier individual behaviors, in particular the practice of physical activity. Career-related decisions are incorporated into the model because the individual's occupation is an endogenous factor that may determine weight status in several different ways. All of the system is estimated by full information maximum likelihood methods; a discrete semi-parametric random effects methodology was used to control for permanent and time-varying unobserved heterogeneity.

Obesity depends highly on an individual's weight status history. Previous obesity is by far the most important factor that explains obesity in a given period. This evidence seems to support the hypothesis that state dependence is more important than observed and unobserved heterogeneity in explaining obesity. For both sexes, this research found that previous obesity increases the probability of current

obesity by close to 70 percentage points. The estimated models predict that once an individual is obese, that state is difficult to leave. This is an important conclusion because the prevention of obesity, especially child or teenage obesity, would clearly be the most efficient strategy.

This research provides evidence that one of the most important strategies for reducing obesity prevalence rate within a generation is the encouragement of physical activity (PA). This is especially true for females, for whom the obesity prevalence rate in 2008 was almost 36% (substantially higher than for their male counterparts). However, not all types of PA are significant in the reduction of the probability of obesity. In the case of women, after controlling for the endogeneity of the PA, only intense physical activity (at least five times per week) was significant in the reduction of the probability of obesity. Intense physical activity implies a significant reduction of 4.5 percentage points in the probability of being obese. In the case of men, intense physical activity significantly reduced significantly the probability of obesity by 3 percentage points.

The model predicts a sizable reduction in adult obesity as a result of a continuous practice of intense physical activity, and this reduction is stronger for women. For example, in a simulated scenario in which all observed women perform intense physical activity continuously throughout their lives, the obesity prevalence rate was 8 percentage points lower than the model predicted with the observed decisions. This is a reduction of almost 30% in the proportion of obese women in the U.S. as predicted by the model. It is, in a sense, an upper boundary of the potential impacts of a policy of PA encouragement. In the case of men, in the simulated scenario in which all observed men perform intense physical activity continuously throughout their lives, the obesity prevalence rate was 4 percentage points lower than the model predicted with the observed decisions.

The beneficial effects of PA seem to remain throughout individuals' lives. In another exercise in this dissertation, I simulate a situation in which all women and men perform intense physical activity while they are in high school. As a result of this change in their behavior, the obesity prevalence rate of adult women (26–31 years old) was reduced significantly by more than 1 percentage point. In the case of adult men there was a significant reduction of 0.5 percentage points in the obesity prevalence derived from intense physical activity during high school.

An important part of this research is testing the role that neighborhood characteristics play in the encouragement of PA. This is done in a framework in which the residential location decisions are explicitely modeled. Modelling the residential location is important because it helps to control for the endogeneity of neighborhood chracteristics and amenities. Using the econometric framework previously described, I find evidence of a small but significant effect of neighborhood amenities on the reduction in obesity prevalence for men and for women. The small magnitude of this effect, in part, is due to the fact that after controlling for the endogeneity of these variables in the physical activity multinomial

model, most of them are not statistically significant. In the jointly estimated the model for women only one neighborhood amenity significantly increases the probability of medium and intense physical activity. This variable is the square mileage of public parks. In the case of men, two neighborhood amenities significantly increase the probability of intense PA: public and fee-required PA-related neighborhood facilities. Nevertheless, in a simulation exercise in which I increase in one standard deviation the availability of several neighborhood amenities, including the ones previously mentioned, the model predicted a reduction of 1 percentage point in the obesity prevalence rate for adult men and women. This reduction is significant in both cases.

The econometric framework that I propose in this research in order to deal with the endogeneity of neighborhood amenities is a contribution to the literature by itself. There have been no previous attempts of modelling the residential choice in the literature on obesity. In future research I plan to explore the implications of not modelling the residential location decision in models like the presented in this research, where the hypothesis of interest is the influence of neighborhood characteristics on obesity.

What determines the probability of obesity? Using the evidence collected from this research I assess what factors contribute to an increase in the probability of obesity and what factors may be the base of good strategies for obesity reduction. In the case of women, from the set of endogenous variables considered in the model, only intense physical activity is a factor that determines significant changes in the probability of obesity. Smoking, fertility and the consumption of fast food meals are not significant factors in the obesity equation estimated in this model. Considering the evidence provided in this research, the hypothesis that smoking has an effect on female obesity is rejected by the model. The model also rejects the hypothesis that having at least one childbirth during the period comprised between two AddHealth waves increases the probability of obesity. One of the limitations of this research is the lack of a better measure for diet or caloric intake. The proxy that I used here was the number of fast food meals during the week. This variable is not significant in the obesity determination equation, and its marginal effect in the determination of the probability of obesity is negligible.

By far, the main factor determining obesity for women is previous obesity. The probability of obesity increases (non-linearly) with age. Conditional on high educational achievements, occupational choice between white/blue collar jobs is not a factor determining obesity. In comparison with college-educated females, other career definitions that involves lower educational achievements increase the probability of obesity. These effects are significant in one of the jointly estimated specifications. Additional exogenous factors increase significantly the probability of obesity. For African-American females, everything else constant, the probability of obesity is higher in comparison with white females. Similarly, married females and women coming from single-mother households have higher probability of being obese as

well. Some individual characteristics such as having parents with high educational achievements are factors that significantly decrease the probability of obesity for female responders of AddHealh.

As in the case of females, lagged obesity is the main factor that explains the probability of obesity for males. In addition, the probability of obesity increases (non-linearly) with age. Two endogenous factors have significant negative effects in the probability of obesity: intense physical activity and smoking. Contrary to the case of women, for men the model supports the hypothesis that smoking is a factor that reduce the probability of obesity. Fast food meals have a positive effect in obesity, but it is not significant. Conditional on high educational achievements blue collar workers have a higher obesity probability, however, this effect is not significant in one of the jointly estimated specifications. On the other hand, conditional on being a white collar worker, low educational achievements increases significantly the probability of obesity. Individual characteristics such as being a first generation immigrant or having parents with high educational achievements reduce significantly the probability of obesity. Other factors such as being married or Hispanic ethnic background increase the probability of obesity.

Evidence from simulation exercises suggests that the effect of a generalized practice of intense PA during the entire period that the women and men are observed causes a very important reduction in the probability of being obese when they are adults between 26 and 31 years old. These reductions are 4 and 8 percentage points for men and women respectively. The encouragement of intense PA when individual are in high school causes a small significant reduction of the obesity prevalence rate for adult women and men. An increase in PA-related amenities could also help in the reduction of obesity prevalence. From simulation-based exercises I observe a small but significant, reduction in the obesity prevalence rate for adult man and women derived from an increase in a set of PA-related amenities.

References

Angeles, G., Guilkey, D. K. and Mroz, T. A. (1998). Purposive program placement and the estimation of family planning program e ects in Tanzania. *Journal of the American Statistical Association,* 93 (443), 884-899.

Arellano, Manuel, and Stephen Bond. (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58(2):277–97.

Becker, Gary, and Kevin Murphy (1988). "A Theory of Rational Addiction." *Journal of Political Economy* 96(4):675–700.

Bhargava, A, (1991). Identification and panel data models with endogenous regressors, *The Review of Economic Studies* 58, 129–140.

Bhargava, Alok, and John Sargan. (1983). "Estimating Dynamic Random Effects Models From Panel Data Covering Short Time Periods." *Econometrica* 51(6):1635–59.

Bhattacharya and Neeraj Sood (2006), Health Insurance and the Obesity Externality, in Kristian Bolin, John Cawley (ed.) *The Economics of Obesity* (Advances in Health Economics and Health Services Research, Volume 17), pp.279-318

Boone-Heinonen Janne, Gordon-Larsen Penny, Guilkey David, Popkin Barry and Jacobs David (2009). Environment and physical activity dynamics: The role of residential self-selection. *Psychology of Sport and Exercise* (2009)

Boone-Heinonen Janne, Guilkey David, Evenson Kelly and Gordon-Larsen Penny (2010). Residential self-selection bias in the estimation of built environment effects on physical activity between adolescence and young adulthood. *International Journal of Behavioral Nutrition and Physical Activity 2010*, 7:70

Boone-Heinonen Janne and Gordon-Larsen Penny (2009). Life stage and sex specificity in relationships between the built and socioeconomic environments and physical activity. *J Epidemiol Community Health* 2009 105064

Colin Cameron A. and Trivedi Pravin K. (2005) *Microeconometrics: Methods and Applications*. Cambridge University Press, New York May 2005

Chaloupka, Frank. (1991). "Rational Addictive Behavior and Cigarette Smoking." *Journal of Political Economy* 99(4):722–42.

Chattopadhyay, Sudip. (2000). The Effectiveness of McFaddens's Nested Logit Model in Valuing Amenity Improvement. *Regional Science and Urban Economics* v30n(1): 23-43.

Chou, S. Y., Grossman, M., & Saffer, H. (2004). An economic analysis of adult obesity: results from the Behavioral Risk Factor Surveillance System. *Journal of Health Economics*, 23(3), 565-587.

Cutler, D. M., Glaeser, E. L., & Shapiro, J. M. (2003). Why have Americans become more obese? *Journal of Economic Perspectives*, 17, 93–118.

Eid Jean, Overman Henry, Pugad Diego, Turner Matthew, (2008). Fat city: Questioning the relationship between urban sprawl and obesity. *Journal of Urban Economics* 63 (2008) 385–404

Ewing, R., Schmid, T., Killingsworth, R., Zlot, A., Raudenbush, S., (2003). Relationship between urban sprawl and physical activity, obesity, and morbidity. *American Journal of Health Promotion* 18(1), 47–57.

Flegal, K. M. (1995).The influence of smoking cessation on the prevalence of overweight in the United States. *N. Engl. J. Med.* 333 (18):1165–70.

Flegal Katherine, Ogden Carroll, Curtin Lester, (2010). Prevalence and Trends in Obesity Among US Adults, 1999-2008. *Clinician's Corner* January 20, 2010—Vol 303, No. 3

French Michael, Norton Edward, Fang Hai, Johana MaClean (2010), Alcohol Consumption and Body Weight. Health Economics 19: 814–832 (2010).

Folmann Nana B., Skovgaard Bossen Kristine, Willaing Ingrid, Sorensen Jan, Sahl Andersen John, Ladelund Steen, Jorgensen Torben (2006), Obesity, Hospital Services use and Costs, in Kristian Bolin, John Cawley (ed.) *The Economics of Obesity* (Advances in Health Economics and Health Services Research, Volume 17), pp.319-332

Friedman, J. (1975). Housing Location and the Supply of Local Public Services. Ph.D. dissertation, Department of Economics, University of California, Berkely, CA.

Gerace, T.A., et al. (1991). Smoking cessation and change in diastolic blood-pressure, body weight, and plasma-lipids. *Prev. Med.* 20(5):602–20.

Giles-Corti, B., Macintyre, S., Clarkson, J.P., Pikora, T., Donovan, R.J.,(2003). Environmental and lifestyle factors associated with overweight and obesity in Perth, Australia. *American Journal of Health Promotion* 18 (1), 93–102.

Gilleskie Donna and Strumpf Koleman, (2005). The Behavioral Dynamics of Youth Smoking, *Journal of Human Resources*, University of Wisconsin Press, vol. 40(4), pages 822-866.

Glaeser, E.L., Kahn, M.E., 2004. Sprawl and urban growth. In: Henderson, V., Thisse, J.-F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. North-Holland, Amsterdam, pp. 2481–2527.

Gordon-Larsen P, Nelson MC, Page P, Popkin BM. (2006) Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics*. 2006;117(2):417-424.

Green, M. S., and G. Harari (1995). A prospective study of the effects of changes in smoking habits on bloodcount, serum lipids and lipoproteins, body weight and blood pressure in occupationally active men: The Israeli CORDIS study. *J. Clin. Epidemiol.* 48(9): 1159–66.

Grossman, M. (1972): .On the Concept of Health Capital and the Demand for Health,. *Journal of Political Economy* 80, 223.255.

Grunberg, N. E., and L. C. Klein. (1998). The relevance of stress and eating to the study of gender and drug use. In National Institute on Drug Abuse Report: Drug addiction research and the health of women, ed. C. L.Weatherington and A. B. Roman. NIH publication no. 98-4290. Rockville, MD: U.S. Dept. of Health and Human Services, National Institutes of Health, National Institute on Drug Abuse.

Gunderson Erica and Abrams Barbara (2000). Epidemiology of Gestational Weight Gain and Body Weight Changes After Pregnancy. *Epidemiologic Reviews* Vol 22 #2

Heckman, J., and Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52, 271-320.

Kaufman, L., and P. J. Rousseeuw. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley.

Keppel KG, Taffel SM (1993). Pregnancy-related weight gain and retention: Implications of the 1990 Institute of Medicine guidelines. *Am J Public Health* 1993, 83:1100-3.

Labeaga, Jose. (1999). "A Double-Hurdle Rational Addiction Model with Heterogeneity: Estimating the Demand for Tobacco." *Journal of Econometrics* 93(1):49–72.

Lakdawalla, D., & Philipson, T. (2002). The growth of obesity and technological change: A theoretical and empirical examination (NBER Working Paper No. 8946).

Lakdawalla, D., Philipson, T. J., & Bhattacharya, J. (2005). Welfare-enhancing technological change and the growth of obesity. *American Economic Review*, 95(2), 253-257.

Lathey Vasudha, Guhathakurta Subhrajit and Aggarwal Rimjhim. (2009). The Impact of Subregional Variations in Urban Sprawl on the Prevalence of Obesity and Related Morbidity. *Journal of Planning Education and Research* 2009 29: 127

Liu, H., T. Mroz, and W. van der Klaauw (2010): "Maternal Employment, Migration, and Child Development," *Journal of Econometrics*, 156(1), 212–228.

McFadden, D. (1978): *Spatial Interaction Theory and Planning Models.* Modeling the Choice of Residential Location, pp. 75–96

Mizoue,T., et al. (1998). Body mass decrease after initial gain following smoking cessation. *Int. J. Epidemiol.* 27(6):984–88.

Moffitt, R., Fitzgerald, J. and Gottschalk P. (1999), "Sample Attrition in Panel Data: The Role of Selection on observables," *Annale d' Economies et de Statistique* 55/56, 129-152.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., et al. (2001). Prevalence of obesity, diabetes, and obesity-related health risk factors. *Journal of the American Medical Association*, 289, 76–79.

Mokdad, A. H., Marks, J. S., Stroup, D. F., & Gerberding, J. L. (2000). Correction: Actual causes of death in the United States. *Journal of the American Medical Association,* 293, 293–294.

Mroz, Thomas. (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.

Mroz, T., Guilkey, D., 1992. Discrete factor approximations for use in simultaneous equation models with both continuous and discrete endogenous variables. Mimeo, Department of Economics, University of North Carolina. Chapel Hill.

Mroz, Tom, and Tim Savage, 2006, The long-term effects of youth unemployment, *Journal of Human Resources* 41, 259–293.

Must, A., Spadano, J., Coakley, E. H., Field, A. E., Colditz, G., and Dietz, W. H. (1999). The disease burden associated with overweight and obesity. *Journal of the American Medical Association*, 282, 1523–1529.

O'Hara, P., et al. (1998). Early and late weight gain following smoking cessation in the lung health study. Am. *J. Epidemiol.* 148(9):821–30.

Ohlin A, Rossner S (1990). Maternal body weight development after pregnancy. *Int J Obes* 1990; 14:159-73.

Papas Mia, Alberg Anthony, Ewing Reid, Helzlsouer Kathy, Gary Tiffany, Ann Klassen (2007). *Epidemiol Rev* 29(1): 129-143

Parker JD, Abrams B (1993). Differences in postpartum weight retention between black and white mothers. *Obstet Gynecol* 1993;81:768-74.

Parsons, G., and M. Kealy (1992): "Randomly Drawn Opportunity Sets in a Random Utility Model ofLake Recreation," *Land Economics*, 68(1), 93–106.

Philipson, Tomas J. and Richard A. Posner. (1999). "The Long-Run Growth in Obesity as a Function of Technological Change." NBER Working Paper No. 7423

Rashad Inas, Grossman Michael (2004). The economics of obesity. *Public Interest* Summer 2004; 156; pp. 104-112

Rashad, I. (2006). Structural estimation of caloric intake, exercise, smoking, and obesity. *The Quarterly Review of Economics and Finance*, 46(2), 268-283.

Rosen, Sherwin. (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82, no. 1 (January): 34–55.

Rossner S, Ohlin A (1995). Pregnancy as a risk factor for obesity: lessons from the Stockholm Pregnancy and Weight Development Study. *Obes Res* 1995; 3 (Suppl 2):267s-75s.

Saelens, B.E., Sallis, J.F., Black, J.B., Chen, D., (2003). Neighborhood based differences in physical activity: An environment scale evaluation. *American Journal of Public Health* 93 (9), 1552–1558.

Thaler Richard (1980). Toward a positive theory of consumer choice. Journal of Economic Behavior and Organization, 1(1):39–60, March 1980.

Train, K., D. McFadden, and M. Ben-Akiva (1987): "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices," *The RAND Journal of Economics*, 18(1), 109–123

US Department of Health and Human Services. (2001). The Surgeon General's call to action to prevent and decrease overweight and obesity. Washington, DC: US Government Printing Office.

Wen Shu, Norton Edward, Guilkey David, Popkin Barry. (2010). Estimation of a Dynamic Model of Weight. NBER Working Paper No. 15864

Wolf, A., & Colditz, G. (1998). Current estimates of the economic cost of obesity in the United States. *Obesity Research*, 6, 97–106.

Wooldrigge Jeffrey (2002), Econometric Analysis if Cross Section and Panel Data. The MIT Press Cambridge Massachusetts.

Yang Zhou, Gilleskie Donna, Norton Edward (2009). Health Insurance, Medical Care, and Health Outcomes A Model of Elderly Health Dynamics. *The Journal of Human Resources* 44-1.

7 Appendices (Omitted in this version of the Paper)