Obesity and Health-Related Decisions: An Empirical Model of the Determinants of Weight Status

Por: Leonardo Fabio Morales, Penny Gordon-Larsen, David Guilkey

Borradores de ECONOMÍA

Núm. 846
2014

Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia

Obesity and Health-Related Decisions: An Empirical Model of the Determinants of Weight Status*

Leonardo Fabio Morales
Banco de la República
lmoralzu@banrep.gov

Penny Gordon-Larsen
Carolina Population Center
pglarsen@unc.edu

David Guilkey
Carolina Population Center
dguilkey@email.unc.edu

Abstract

Using Add Health, a very comprehensive longitudinal data set of teenagers and young adults in the United States, we estimate a structural dynamic model of the determinants of obesity. In addition to including many of the well-recognized endogenous factors mentioned in the literature as obesity determinants, i.e., physical activity, smoking, a proxy for food consumption, and childbearing, we also model the residential location as a choice variable, relevant to the young-to middle-aged adult, as a major component. This allows us to control for an individual's self-selection into communities which possess the types of amenities in the built environment which in turn affect their behaviors such as physical activity and fast food consumption. We 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 a semi-parametric full information log-likelihood method that allows for a general pattern of correlation in the errors across equations. Simulations are then used to allow us to quantify the effects of these endogenous factors on the probability of obesity. A key finding is that controlling for residential self-selection has important substantive implications. To our knowledge, this has not been yet documented within a full information maximum likelihood framework.

Keywords: Health Production, Public Health, Urban Analysis

JEL Codes: I12, I14, R52

*This work is preliminary and incomplete. The opinions expressed here are those of the authors' and not of the Banco de la República de Colombia nor of its Board. All errors are authorss' responsibility. We thank participants of the UNC Chapel Hill Applied Microeconomics Workshop for their helpful comments and advice.

1 Introduction

In recent years health economists 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 such as Type II diabetes, heart disease, and hypertension (Mokdad et al., 2005; 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. In 2008 in the US, 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 relative importance of factors that cause obesity and their contribution to the remarkably high obesity prevalence in the US. Several studies have focused on the effect that the relative prices of calories and physical activity have on the determination of weight status. More in line with this research are efforts that have been made to find causal links between individual choices such as smoking, physical activity, and diet and obesity . Several authors have noted that these individual choices are endogenous (French et al., 2010, Grossman and Saffer 2004; Rashad 2006; Wen et al., 2010). Another line of research has explored the impact that the environment in which individuals perform their daily activities has on obesity. Recent literature in epidemiology, urban economics, and planning has focused on the role of built environments in increasing energy consumption and decreasing energy expenditure. In other words, neighborhoods may affect opportunities for exercise and healthy diet behaviors, and thereby have an impact on obesity. (Ewing, Schmid, Killingsworth, Zlot, and Raudenbush, 2003; Giles-Corti, Macintyre, Clarkson, Pikora, and Donovan, 2003; Glaeser and Kahn, 2004; Lathey, Guhathakurta, and Aggarwal, 2009; Saelens, Sallis, Black, and Chen 2003, Boone-Heinonen and Gordon-Larsen, 2009; Gordon-Larsen, Nelson, Page, and Popkin, 2009).

In this paper we 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 unobserved heterogeneity. 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 semi parametric full information maximum likelihood methods.

In addition to taking into account lifestyle choices (smoking, physical activity, etc.) that have been linked to obesity in the literature, our research also models a choice variable of particular relevance to the young- to middle-aged adult, the residential location decision. This decision determines the characteristics and resources of neighborhoods in which individuals live. These characteristics and resources may encourage individuals to increase their energy expenditure levels (e.g., through amenities that encourage physical activity) or energy intake (e.g., through availability of fast food restaurants). By modeling the residential location decision, we are able to control for the potential endogeneity of neighborhood characteristics in the decision to perform any sort of physical activity We show that modelling residential choice decisions is important and that the effect of neighborhood characteristics on obesity are biased if researchers ignore the fact that individuals self-select themselves into neighborhoods.

Using the estimated model, we use simulations to quantify the contribution of the endogenous factors on the probability of an individual being obese. We also use simulations to trace the pathways through which neighborhood amenities impact on individuals' endogenous lifestyle decisions and then ultimately impact obesity through these intermediate outcomes., Thus, our research contributes to the recent debate about the influence of the built environment on the propensity to become obese.

We find evidence of a significant reduction in the obesity prevalence for adult females and males derived from a hypothetical situation in which they perform intense physical activity when they were high school students. We also find evidence that a generalized, continuous practice of intense physical activity would produce big drops in adult obesity prevalence. In addition, we find evidence of an important reduction in obesity prevalence, especially for men, caused by a reduction in the consumption of fast food. We also test if neighborhood amenities have a significant impact upon physical activity levels. After controlling for the endogeneity of neighborhood amenities, we find a small but significant effect of a set of physical activity related facilities on physical activity behaviors and obesity. For women we find a negative and significant effect of street connectivity, a measure of urban density and walkability, on obesity prevalence.

2 Data

The main source of data used in this study is The National Study of Adolescent Health (Add Health). A key feature of the dataset 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 describing the cohort and the contextual information available in Add Health and our definition of neighborhoods.

2.1 The Add Health Study

Add Health is a prospective cohort study of 20,745 adolescents, representative of the U.S. school-based population in grades 7 to 12 (mean age 15.4 years in 1994-95) and followed over four examinations into adulthood when respondents were 24-34 years of age. The current study uses data from Wave I (n=20,747); II (n=14,738; mean age 16.3 years in 1996), Wave III (n=15,197; mean age 21.7 years in 2002), and Wave IV (n = 15,701; mean age 28.3 years in 2009) of Add Health. The study population was obtained through a systematic random sample of 80 high schools and 52 middle schools in the United States, stratified to ensure that the schools were representative of U.S. schools grades 7 through 12 with respect to region, urbanicity, school type, percentage of white students, and school size from 80 communities across urban, suburban, and rural areas. Add Health included a core sample plus subsamples of selected minority and other groupings collected under protocols approved by the Institutional Review Board at the University of North Carolina at Chapel Hill. The survey design and sampling frame have been described elsewhere (Miller et al,2014; Harrys, 2010).

An important component of Add Health is the collection of anthropometric data, as well as extensive information on respondent's health-related behaviors. Of particular relevance is the collection of data on the activities that respondents perform during their "active" leisure time, as physical activity is a key determinant of obesity. Furthermore, physical activity is the most malleable component of energy balance and is a strong policy target in efforts to reduce and prevent obesity.

2.2 Contextual Information and Neighborhood Characteristics

Add Health contains a very large amount of contextual information. This information describes a comprehensive set of characteristics of the environments in which Add Health respondents live and is available at small geographic units, something that is not very common in public versions of longitudinal datasets. Many variables are available at the Census tract level, and some are generated for even smaller geographic areas. An important subset of contextual variables, some of which are used in this study, was generated to describe characteristics of the area within a specified radial distance (1, 3, 5, and 8 km) from each respondent's residential address at each survey wave.

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 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[1]".Some other contextual variables used in the present study were generated using the previously described residential buffers. In this case we selected a 5k buffer because the best model fit was obtained using this radius.

In this research we are able to identify each respondent's neighborhood at each survey wave. Therefore, we know attributes of the neighborhoods where respondents are located. Although contextual information is available for all waves of the Add Health study, some variables are not available for all respondents in the estimation samples at all waves. In order to be able to use the contextual information despite this missing values problem, we 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.

3 Theoretical Motivation

In the dynamic framework in which the individuals are modelled in this research, they are allowed to make choices about their residential location. Similarly, they make decisions about food consumption, marriage, fertility and smoking and also decide the amount of leisure physical activity they engage in. As a result of all these choices and the relationships among them, the weight of the individual is produced as a health outcome. The existence of a causal relationship between weight and food consumption, and between weight and physical activity, is well known. Smoking has been widely proven to have a negative impact on an individual's weight (Mizoue et al., 1998; O'Hara et al., 1998). In addition, for women who have given birth, obesity may be due to weight retention after delivery (Gunderson Abrams, 2000; Rossner and Ohlin, 1995).

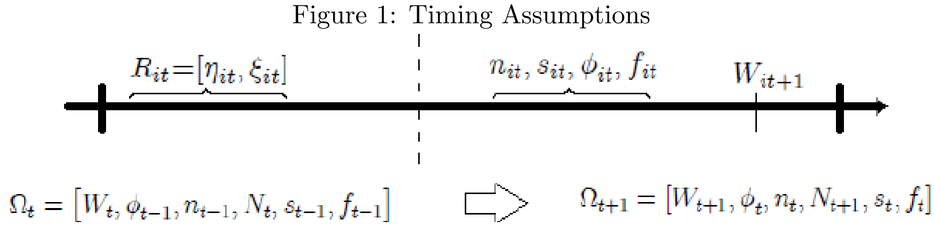
An important feature of this model is that it is dynamic in the sense that previous behaviors influence current decisions. The theoretical justification for this comes from the traditional theory of rational addiction (Becker and Murphy, 1988) wherethe utility of an addictive good is influenced by previous consumption behavior. The rational addiction framework has been traditionally used to model risky behaviors such 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).

[1] 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

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 level 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 decision in the period, the residential location 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}, whether or not to smoke s_{it}, and level 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 assumption in this time framework is important for identification purposes: shock prices that affect the choices of $[n_{it}, s_{it}, \phi_{it}, f_{it}]$ occurring within the second period occur after choices made during the first period. Therefore, individuals learn about these shocks after they have made the residential location decisions. These shock prices at the beginning of the second period could be interpreted as new information that appears between the first and second periods. The residential choice is a major decision that influences the residential location, and thus exposure to different distributions of prices, for any given individual. These shock prices are stacked in the vector $\Psi_t = \left[p_{\eta_t}, p_{s_t}, p_{\phi_t}, p_{f_t}\right]$.

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 influence 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 individual i in period t can by represented as:

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

At any period t, the objective of the individual in this model is to maximize the expected present discounted value of remaining lifetime utility.

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

Subject to standard time and budget constraints. Demand functions for the choice variables result from the solution to this optimization problem. Substituting these demand functions into a health production function that explains the individuals weight, yields an estimable expression for the individuals' weight $W_{it+1} = W(W_{it}, a_{it}, s_{it}, n_{it}, f_{it})$. Approximations for all these equations are estimated in the empirical model.

4 Empirical Model

The weight outcome is a function of the individual's choice variables, and the endogeneity of these choices must 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 on the final indirect effect that they have on the individual's weight status; furthermore, unobservables affecting each one of these behaviors may be correlated. 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.

The timing strategy described in Figure 1 assumes that when individuals make decisions, they consider all the information available at the moment of the decision. The available information at the beginning of the period is composed of previous choices and state variables. This sequence implies that the model is dynamic and part of the identification of the model will be based on this dynamic nature. In addition, the empirical model also includes 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.

4.1 Error Structure

Observed factors do not explain all variation in the outcomes that are modeled. Unobserved characteristics also influence each one of these behaviors and these unobserved characteristics may be correlated across equations. The correlation patterns 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. We denote each one of the equations in the system by $e = \{1, 2, ..., 6\}$, 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 with unobserved characteristics related to differences in 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 differentially affects their behavior. If these unobserved heterogeneity terms are correlated across outcomes, then bias is introduced if they are not taken into account. Whether or not there is correlation is a testable assumption.

In order to estimate the parameters of the unobserved heterogeneity distribution, we use a semi-parametric discrete factor approximation method. The Discrete Random Method is more general than other methodologies that assume an arbitrary distribution for the unobserved heterogeneity (Heckman and 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, and Guilkey, 1998, Guilkey and Lance, 2014). The joint distribution of the unobserved effects is modeled as a multivariate discrete distribution with multiple points of support and is estimated jointly with all other parameters of the model. Using Monte Carlo experiments, it has been shown that when the real underlying distribution for unobserved heterogeneity is normal, the discrete factor random effects (DFRE) method performs almost as well as methods that assume normality. However, when the true error distribution is not normal, the DFM is substantially better than methods that incorrectly assume normality in terms of the precision and accuracy of the estimators (Mroz, 1999; Mroz and Guilkey, 1992, Guilkey and Lance, 2014).

4.2 Empirical Equations

4.2.1 Residential Location Decisions

Mixed Logit Specification

We assume that place of residence is exogenous to the individual at waves I and II since almost all respondents are still living with their parents and we model choice of residence at waves III and IV. We feel that it is necessary to model residential location since individuals may chose a location because of the amenities available that could in turn affect physical activity and other behaviors and the endogeneity of these amenities must be controlled.

The specification of the equation for residential decisions is modeled as a mixed logit, which is a conditional logit augmented with non-alternative varying characteristics. The specification of 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. 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:

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

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 Add Health 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:

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

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 which makes the estimation intractable unless we limit the number of alternatives from which the individuals are allowed to choose. In this study we use a sampling technique to solve with this problem. The vector Z_{itj} in Equation 5.22 collects a set of location-specific variables or amenities in location j. In the specification, we also include individual-specific regressors similar to the ones included in the vector of exogenous individual characteristics X_{it}, and previous 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. Note that we 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) for each neighborhood in the individual's choice set.

The method we use to deal with the intractable choice set was a random sampling of the individual's neighborhood choices. Under some minimal conditions this technique has been proven to provide consistent estimators using a subset of the choice set (McFadden, 1978). In this paper the choice set K^R is partitioned into sets $\{C_1, C_2, ..., C_L\}$. The choice subset will be formed by taking chosen neighborhood k from partition set C_k with one randomly selected alternative from each remaining partition sets. The previously described sampling procedure requires the partition of the choice set K^R. After this partition is implemented, the random sampling is performed within each of the partition sets. In this study we use partitions defined as types of neighborhoods. We use a non-hierarchical cluster procedure to form four clusters based on different neighborhood characteristics. Each cluster defines a different type of neighborhood based on the characteristics and amenities that the neighborhood has.

4.2.2 Simultaneous Choices and Final Health Outcome

Physical activity is directly translated into energy expenditure and is thus a metabolic determinant of weight. In this paper, physical activity is modeled as a categorical variable that describes the frequency with which several physical activities 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 implies that smoking, diet and fertility decisions do not contemporaneously affect physical activity.

As already mentioned previously, smoking is treated as an endogenous variable. 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. 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, we use a logit model that specifies whether the women had at least one child during period t. Marriage is a choice variable that is highly related to childbirth. There is no information in Add Health that allows us to construct a measure of caloric intake. In order to deal with this issue, we use as a proxy variable the frequency of fast-food meals in a week. A high frequency of fast-food meals would be consistent with a high amount of calories and a signal of poor-quality diet because fast food is usually cheap and calorie-dense. The positive significant relationship between weight and fast food has been noted previously in the literature (Chou, Grossman, and Saffer, 2004). The following system of equation represents the set of inputs that contemporaneously determined the individual's weight status at the end of the period.

$$\ln\left[\frac{P\left(A_{it}=k^A\right)}{P\left(A_{it}=1\right)}\right] = X_{it}\gamma^A + Z_{it}\beta^A + C_{it}\delta^A + \Omega_{it}\phi^A + I_{it}\theta^A + \mu_i^A + \upsilon_{it}^A \quad \text{(5.3: Physical Activity)}$$

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

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

$$F_{it} = X_{it}\gamma_{kf}^F + Z_{it}\beta_{kf}^F + C_{it}\delta_{kf}^F + \Omega_{it}\phi_{kf}^F + I_{it}\theta_{kf}^F + \mu_i^F + \upsilon_{it}^F + u_{it} \quad \text{(5.6: Fast food meals)}$$

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

Matrix X_{it} includes individual demographic and socioeconomic exogenous characteristics. Matrix Z_{it} includes amenities of the individual's place of residence. C_{it} is a matrix of career dummies, where career is a set of dummies that indicates the individual's main occupation. In this paper occupation is understood to be a combination of educational achievement and occupational position (students, white/ blue collar workers, etc.). Finally, 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 of exogenous variables that impact these simultaneous behaviors. The error structure is the same as in previous equations in that it allows for unobserved constant and time- varying unobserved heterogeneity.

4.2.3 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-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 behaviors. 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

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

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 υ_{it}^{w} represents time varying unobserved individual characteristics.

4.3 Initial Conditions and Identification Issues

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 even under fairly weak conditions, the system can be 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 an 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. In addition, use of non-linear models of unobserved heterogeneity contributes to identification as well (see Guilkey and

Lance , 2014). 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, we incorporate exclusion restrictions in the equations for lifestyle decisions. These are neighborhood amenities and local prices that determine decisions related to engaging in physical activity and other lifestyle decisions, but presumably they do not have a direct impact on weight status. The set of exclusion restrictions in our system includes a factor score of physical activity- related amenities, non-physical activity related amenities within a 5km residential buffer, local crime measured by violent arrests per 10,000 inhabitants, use of any method of contraception, and community-level prices from the Council for Community and Economic Research (C2ER): cigarette, groceries, and junk food prices per 2005 dollars. All of these variables are assumed to have a direct impact on life style choices, and an indirect impact on weight status only through their effect on lifestyle choices. Empirically we verified that all these variables are not jointly significant in the structural equation, using a log-likelihood ratio test[2].

The model is dynamic, which evokes another standard concern: the initial condition problem for 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. 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.

[2]The way we test joint significance of exclusion restrictions in the structural equation is by comparing the likelihood function of two estimations: one that excludes these variables from the structural obesity equation, and a second that that includes them in the equation. With this information we test the hypothesis using a standard likelihood ratio test. It is worth mentioning that the nonlinearity of the model allows identification without the necessity of exclusion restrictions, therefore, this is what allows us to test the over-identifying restrictions in the model.

4.4 Estimation

The methodology for the estimation of all equations in section 5, including the initial condition equations, is based on full information semi parametric maximum likelihood methods (FIML). 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.

As explained in section 4.1, in this study we 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:

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

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

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

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:

$$
L\left(\Theta\right) = \prod_{i=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k \left[\prod_{t=2}^{4} \sum_{l=1}^{L} \psi_l L\left(\Theta | \mu_k, \upsilon_{l,t}\right) \right] \right\}
$$

where $L\left(\Theta | \mu_k, \upsilon_{l,t}\right)$ is the individual and period specific contribution to the likelihood conditional on the values for the unobserved heterogeneity parameters $\mu_k, \upsilon_{l,t}$. The whole set of parameters estimated in the model is:

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

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

Finally, we control for the attrition in the panel. We use a methodology based on inverse probability weighting (Horowitz and 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 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.

5 Results

In this section we present the estimation results for the system of equations and we present results from simulation-based experiments using parametric bootstrap methodologies. We begin by presenting summary statistics for the data. For a convenient presentation of the results, 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 choice of residential location are in Appendix B.

5.1 Summary Statistics and Sample Description

The initial sample of the Add Health study includes more than 20,745 individual respondents. After merging all waves to construct the panel of individuals, a large fraction of respondents is lost to follow-up. The final sample size of individuals was reduced to 4,400 women and 3,660 men, observed in four waves with complete data. In our sample we have 31% (36%) women (men) who report regular consumption of cigarettes at some point during the span they were interviewed between wave 2 and wave 4. In this sample 25% (35%) of women (men) reported they performed high levels of physical activity between wave 2 and wave 4. The average number of fast food meals reported consumed per week between wave 2 and wave 4 is 2.13 and 2.55 for men and women respectively. In addition, an average prevalence of 23% and 24% of obesity was observed between wave 2 and wave 4 for men and women respectively. Complete summary statistics of the variables used in the estimations of the lifestyle and the weight status equations is presented in Table 1. Additional details about the construction of variables are presented in the Data Appendix.

5.2 Estimations Results of the Model

Obesity Equation

The focus of the model is the estimation of the determinants of obesity. Table 2 includes the results of the estimation of this equation for men and women. Two different specifications are presented, in four columns. Columns 1 and 2 contains the results of an individual logit with no endogeneity control (unobserved heterogeneity is not modeled in any way). In specifications 3 and 4, the obesity equation is jointly estimated with all other equations in the system and with initial conditions. In this specification we include four (women) or three (men) permanent and three time-varying unobserved heterogeneity parameters. A table with the estimated unobserved heterogeneity parameters is presented in Appendix B. Inclusion of additional unobserved heterogeneity parameters (permanent and time-varying) did not improve the performance of the models in terms of a significant increment of the log likelihood function.

One strong result from the estimation of the weight status equation is the persistence of obesity. Note from columns [3] and [4], in the jointly estimated models, the probability of being currently obese increased by 66 and 63 percentage points as a function of past obesity for women and men, respectively. Given this level of persistence, the importance of obesity prevention early in the lifecycle is crucial. The most important determinant of obesity in the future, by far, is a high BMI today. Individuals who become obese as teens or during the entry to the adult years are likely to carry the burden of obesity throughout their lives.

Physical activity (PA) significantly reduces the probability of being obese, for women and men, but only for those who perform the highest amount of physical activity (at least five episodes in a week). There is a significant reduction of almost 4 percentage points for women and almost 6 percentage points for men in the probability of obesity for those who perform at least 5 episodes of weekly PA. The magnitude of the effect of high levels of PA is higher in the individual logit estimation for women and a bit lower for men, nevertheless, this marginal effect for the probability of being obese remained negative and statistically significant in jointly estimated models. Clearly, physical activity could be a very important tool for reducing or preventing obesity in women and men. Simulations derived from the estimated model show that the obesity prevalence rate would be greatly reduced with a generalized implementation of this practice. These simulations will be discussed in the next section.

The dummy variable representing fertility has a positive but non-significant effect in specifications [1] and [3]. Cigarette smoking is not statistically significant for women but it is for men, with a 3

Table 1: Summary Statistics

Variable	Women			Men		
	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
Obesity	11265	0.232	0.422	10500	0.240	0.427
>=1 Childbirths	11265	0.248	0.432			
Smoker	11265	0.312	0.463	10500	0.363	0.481
PA 3 or 4 times per week	11265	0.486	0.500	10500	0.490	0.500
PA 5+ times per week	11265	0.247	0.432	10500	0.356	0.479
# Fast Food meals	11265	2.137	2.010	10500	2.553	2.263
Family Income ($16K-$30k)	11265	0.131	0.338	10500	0.121	0.326
Family Income ($30K-$50k)	11265	0.169	0.375	10500	0.169	0.374
Family Income ($50K- $100K)	11265	0.288	0.453	10500	0.307	0.461
Family Income (+$100K)	11265	0.155	0.362	10500	0.183	0.386
Family Income Missing	11265	0.107	0.309	10500	0.103	0.304
High Education/ White Collars	11265	0.127	0.333	10500	0.096	0.294
High Education/ Blue Collars	11265	0.124	0.330	10500	0.152	0.359
Med-Low Education/White Collars	11265	0.112	0.316	10500	0.091	0.288
Med-Low Education/Blue Collars	11265	0.095	0.294	10500	0.171	0.376
Med-Low Education/Not working	11265	0.090	0.286	10500	0.066	0.248
Age	11265	21.841	5.147	10500	22.070	5.178
Age2	11265	503.505	228.536	10500	513.903	232.250
African American	11265	0.222	0.416	10500	0.178	0.382
Asian	11265	0.052	0.221	10500	0.063	0.243
Hispanic	11265	0.132	0.339	10500	0.145	0.352
1st generation immigrant	11265	0.042	0.200	10500	0.045	0.208
2nd generation immigrant	11265	0.059	0.235	10500	0.065	0.246
Married	11265	0.200	0.400	10500	0.160	0.366
Cohabitating	11265	0.125	0.331	10500	0.121	0.326
Divorced or separated	11265	0.085	0.279	10500	0.079	0.270
Living with Parents	11265	0.511	0.500	10500	0.556	0.497
No. Children younger than 6	11265	0.339	0.667	10500	0.193	0.542
No. Children older than 6	11265	0.122	0.432	10500	0.055	0.311
Family size	11265	3.724	1.788	10500	3.610	1.677
Initial H/H: Step parents	11265	0.142	0.350	10500	0.141	0.348
Initial H/H: Single Father	11265	0.018	0.132	10500	0.029	0.168
Initial H/H: Step Mother	11265	0.193	0.394	10500	0.171	0.376
Initial H/H: Non Parents	11265	0.054	0.227	10500	0.040	0.195
Parents Education: High School	11265	0.241	0.428	10500	0.231	0.422
Parents Education: Some College	11265	0.266	0.442	10500	0.275	0.446
Parents Education: Bachelor	11265	0.182	0.386	10500	0.199	0.400
Parents Education: +Bachelor	11265	0.147	0.354	10500	0.148	0.355
Parents Education: Missing	11265	0.056	0.231	10500	0.049	0.215
Dummy third wave	11265	0.333	0.471	10500	0.333	0.471
Dummy Fourth wave	11265	0.333	0.471	10500	0.333	0.471
Beta Street Conectivity Index (B)	11265	1.432	0.125	10500	1.426	0.124
Miles^2 of parks	11265	0.645	2.497	10500	0.662	2.513
PA Related Amenities Factor Index (Pc1)	11265	-0.456	1.488	10500	-0.479	1.397
Violent Arrest by 10000 inhabitants	11265	8.726	4.612	10500	8.583	4.492
Non PA related Amenities 5km buffers	11265	0.222	0.725	10500	0.209	0.689

Table 2: Weight Status Estimation Results

Variable	[1]: Women				[2]: Men				[3]: Women				[4]: Men			
	Logit								Jointly estimated and Sampling of the Neighborhood Choice Set[1]							
	Coef	S. D.		Mfx	Coef	S. D.		Mfx	Coef	S. D.		Mfx	Coef	S. D.		Mfx
Constant	-5.81	1.17	***		-3.00	1.18	**		-6.57	1.30	***		-3.81	1.36	***	
Beta Street Conectivity Index (B)	0.27	0.44		0.028	-0.44	0.47		-0.046	0.27	0.44		0.027	-0.48	0.51		-0.053
B x {2nd Neighborhood Median Income Quartile}	-0.71	0.64		-0.062	0.80	0.68		0.112	-0.69	0.64		-0.058	0.98	0.72		0.144
B x {3rd Neighborhood Median Income Quartile}	-1.55	0.69	**	-0.125	0.76	0.74		0.106	-1.55	0.69	**	-0.120	0.96	0.77		0.141
B x {4th Neighborhood Median Income Quartile}	-0.32	0.67		-0.030	0.57	0.68		0.076	-0.29	0.67		-0.026	0.81	0.72		0.114
Obese previous pd	3.74	0.09	***	0.673	3.84	0.09	***	0.703	3.71	0.10	***	0.660	3.37	0.20	***	0.627
>=1 Childbirths	0.11	0.12		0.011					0.08	0.12		0.008				
Smoker	0.00	0.07		0.000	-0.21	0.07	***	-0.024	0.11	0.08		0.011	-0.25	0.08	***	-0.030
PA 3 or 4 times per week	-0.09	0.07		-0.009	-0.14	0.09		-0.016	-0.08	0.07		-0.008	-0.15	0.09		-0.018
PA 5+ times per week	-0.43	0.10	***	-0.040	-0.48	0.11	***	-0.053	-0.42	0.11	***	-0.037	-0.48	0.12	***	-0.056
# Fast Food meals	-0.01	0.01		-0.001	0.01	0.01		0.001	0.02	0.03		0.002	0.06	0.03	**	0.007
Family Income ($16K-$30k)	0.21	0.11	**	0.021	0.00	0.14		0.000	0.22	0.11	**	0.022	-0.01	0.14		-0.001
Family Income ($30K-$50k)	0.12	0.10		0.012	0.00	0.13		-0.001	0.13	0.11		0.013	0.03	0.13		0.004
Family Income ($50K- $100K)	0.06	0.10		0.006	0.08	0.12		0.010	0.09	0.10		0.009	0.13	0.13		0.016
Family Income (+$100K)	-0.16	0.13		-0.016	0.08	0.14		0.009	-0.13	0.13		-0.012	0.14	0.14		0.018
Family Income Missing	0.06	0.13		0.006	-0.22	0.15		-0.025	0.08	0.13		0.008	-0.17	0.16		-0.021
High Education/ White Collars	-0.01	0.13		-0.001	-0.09	0.15		-0.010	0.00	0.13		0.000	-0.07	0.16		-0.008
High Education/ Blue Collars	0.01	0.12		0.001	0.17	0.13		0.021	0.01	0.12		0.001	0.17	0.14		0.021
Med-Low Education/White Collars	0.38	0.12	***	0.040	0.49	0.14	***	0.063	0.38	0.12	***	0.039	0.50	0.15	***	0.066
Med-Low Education/Blue Collars	0.29	0.13	**	0.030	0.07	0.13		0.009	0.29	0.13	**	0.029	0.06	0.14		0.007
Med-Low Education/Not working	0.24	0.13	*	0.025	0.04	0.16		0.005	0.26	0.13	*	0.026	0.00	0.17		0.000
Age	0.20	0.09	**	0.022	0.11	0.09		0.014	0.20	0.09	**	0.021	0.08	0.09		0.010
Age^2	0.00	0.00		0.000	0.00	0.00		0.000	0.00	0.00		0.000	0.00	0.00		0.000
African American	0.44	0.08	***	0.046	-0.02	0.09		-0.003	0.44	0.08	***	0.044	-0.10	0.10		-0.012
Asian	-0.03	0.17		-0.003	-0.13	0.14		-0.015	-0.02	0.17		-0.002	-0.06	0.16		-0.007
Hispanic	0.25	0.10	***	0.026	0.32	0.09	***	0.039	0.26	0.10	***	0.026	0.38	0.11	***	0.049
1st generation immigrant	-0.46	0.19	**	-0.042	-0.32	0.16	**	-0.035	-0.45	0.19	**	-0.039	-0.35	0.19	*	-0.040
2nd generation immigrant	-0.38	0.15	**	-0.035	0.26	0.14	*	0.032	-0.37	0.15	**	-0.032	0.29	0.15	*	0.038
Married	0.20	0.10	**	0.020	0.35	0.11	***	0.043	0.22	0.10	**	0.022	0.42	0.12	***	0.053
Cohabitating	0.11	0.11		0.011	0.16	0.11		0.019	0.12	0.11		0.012	0.18	0.11		0.023
Divorced or separated	0.05	0.10		0.005	0.17	0.12		0.020	0.07	0.11		0.006	0.20	0.12		0.025
Living with Parents	0.11	0.10		0.011	-0.11	0.10		-0.013	0.11	0.10		0.011	-0.12	0.10		-0.014
No. Children younger than 6	0.02	0.07		0.002	0.03	0.07		0.004	0.02	0.07		0.002	0.04	0.07		0.005
No. Children older than 6	-0.16	0.08	**	-0.015	-0.08	0.11		-0.009	-0.15	0.08	*	-0.014	-0.06	0.11		-0.007
Family size	0.06	0.02	**	0.006	0.02	0.02		0.003	0.06	0.02	**	0.005	0.02	0.02		0.003
Initial H/H: Step parents	-0.19	0.09	**	-0.018	0.00	0.09		0.000	-0.22	0.09	**	-0.020	0.03	0.10		0.004
Initial H/H: Single Father	-0.11	0.21		-0.011	-0.41	0.19	**	-0.044	-0.10	0.21		-0.009	-0.49	0.21	**	-0.054
Initial H/H: Step Mother	0.12	0.08		0.012	0.04	0.08		0.004	0.11	0.08		0.011	0.03	0.10		0.003
Initial H/H: Non Parents	-0.15	0.14		-0.014	0.28	0.16	*	0.034	-0.18	0.14		-0.016	0.30	0.19		0.038
Parents Education: High School	0.00	0.10		0.000	0.26	0.11	**	0.031	-0.01	0.10		0.000	0.32	0.13	**	0.040
Parents Education: Some College	-0.12	0.10		-0.012	0.05	0.11		0.006	-0.11	0.10		-0.010	0.12	0.13		0.015
Parents Education: Bachelor	-0.33	0.12	***	-0.031	0.14	0.12		0.017	-0.30	0.12	**	-0.027	0.26	0.14	*	0.033
Parents Education: +Bachelor	-0.35	0.13	***	-0.033	-0.18	0.13		-0.020	-0.32	0.13	**	-0.029	-0.14	0.15		-0.017
Parents Education: Missing	0.19	0.15		0.020	0.20	0.17		0.024	0.22	0.15		0.022	0.27	0.19		0.035
Dummy third wave	0.55	0.17	***	0.055	0.14	0.16		0.017	0.61	0.18	***	0.059	0.39	0.19	**	0.048
Dummy Fourth wave	0.60	0.24	**	0.062	0.49	0.23	**	0.060	0.67	0.25	***	0.068	0.88	0.27	***	0.115
2nd Quartile of Neighborhood Median Income	1.09	0.93		0.126	-1.21	1.00		-0.125	1.00	0.94		0.109	-1.62	1.06		-0.169
3rd Quartile of Neighborhood Median Income	2.08	0.99	**	0.268	-1.23	1.07		-0.130	1.95	1.00	*	0.240	-1.79	1.13		-0.188
4th Quartile of Neighborhood Median Income	0.23	0.96		0.024	-0.87	0.98		-0.098	0.04	0.99		0.003	-1.77	1.05	*	-0.195

Notes: *** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

Quantiles of Neigborhood income were generated using the median household income in the Census Tract. All prices are in 2005 dollars

[1]The definition of the neighborhood-type clusters is based on a Ward cluster procedure with 4 categories

percentage point reduction in the probability of obesity in both the individual logit specification and the jointly estimated model. In our preferred specification (jointly estimated model) there is a positive effect of the number of fast food meals per week for men only. However, from the simulations we show in next section that this variable was an important factor in reducing obesity in both sexes. In comparison with students, women with less than college in any occupational category have higher probability of obesity. However, there is a statistically significant positive effect on obesity for men with less than college education and for men with white collar jobs.

In this paper we test an additional hypothesis: whether neighborhood amenities have an impact on the probability of obesity. Most of the amenities we considered are allowed to have an indirect effect in the obesity equation through the physical activity pathway. Nevertheless, we consider the direct effect of one amenity in the obesity equation, the beta street connectivity index. Higher values of this index reflect more densely connected street networks, which are typically found in areas with greater support for walking and increased development and therefore destinations for shopping, recreation and so forth. We interacted this walkability indicator with neighborhood-income quartile dummies. We use this variable as a proxy for walking, which was not ascertained at all survey waves of Add Health. As readers can see in column [3], in our preferred specification, street connectivity is associated with a decreased probability of obesity for women living in neighborhoods with populations in the third quartile of neighborhood income (on average). We provide a deeper interpretation of this effect in simulations we perform in the next section.

Some exogenous variables, such as the dummy variable for married, dummy variables for African American (just for women) and Hispanic, for example, significantly increase the probability of obesity. These variables remained significant in the jointly estimated model for women and men. Other exogenous variables like first generation immigrant status and high levels of parental education (just for women) significantly reduce the probability of obesity in the individual logit estimation and the jointly estimated model as well.

Input Equations Estimation Results

In this subsection we present the estimation results of the endogenous inputs that are contemporaneously included in the obesity equation: physical activity, smoking, frequency of fast-food meals, and fertility (just for women). For all these equations we estimate two specifications in this paper, independently estimated multinomial logits or linear models, and jointly estimated specifications with all other equations in the system. In jointly estimated models we used random sampling of the choice set of neighborhoods for the estimation of the residential location equation. For ease of presentation, we present the results of our preferred specification (jointly estimated model) in Table 3, results of models independently estimated are presented in an online appendix of the paper.

Physical Activity

Table 3 shows the results of the physical activity equation estimation. The equation for physical activity (PA) is specified as a multinomial model with three categories associated with different levels of PA: two times per week or less (1: Low [referent]), 3 to 4 times per week (2: Moderated), and at least five or more times per week (3: Intense).

Previous obesity significantly reduces the probability of performing medium PA levels for men and high PA levels for women in the subsequent period. For instance, men who were obese in the previous period have a 5 percentage point decrease in the probability of medium PA levels. Previous childbearing has a positive effect (3 percentage points) on the probability of high levels of PA for women. In the jointly estimated model for men and women, cigarette smoking has a statistically significant negative effect on the probability of engaging in high levels of PA, with a reduction of 4 and 5 percentage points in the models for women and men, respectively. In general, past PA is highly predictive of current PA, especially at high levels of PA. In our preferred specification, the probability of high PA levels increases by 6 percentage points for men who performed medium PA (3 to 4 times per week) in the previous period (and by 11 percentage points for women). The estimated reduction in the probability of obesity is even greater at high PA per week (31 percentage points for men; 19 for women). Further, we find a statistically significant reduction in the probability of medium and high levels of PA with each fast-food meal reported per week among women. We also found that participants who were currently attending school had a significant reduction in the probability of high levels of PA relative to almost all other careers (educational-occupational combinations).

The probability of high PA decreases non-linearly and significantly with age in both genders. Some demographic factors significantly reduce the probability of PA as well. For example, the dummy variable for African American is associated with a significant reduction of more than 4 percentage points in the probability of high levels of PA in the specification for women. Having small children at home (younger than 6) also reduces the probability of PA in the specification for women. The probability of high PA levels decreases significantly (more than 5 percentage points) for married and cohabitating men . The negative effect of marriage is not significant for women in the jointly estimated model, but the negative effect of cohabitating is statistically significant.

To summarize, some of the most remarkable features of the PA estimation are that respondents of Add Health are significantly more likely to perform high levels of PA in the current period if they did not smoke during the previous period, but they engaged in medium or high levels of PA. In addition, more frequent fast-food meals in the previous period (in the case of women) are associated with a significant reduction in the probability of performing medium and high 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 an impact on the probability of obesity. One of the ways we tested this hypothesis was by evaluating the effect that a set of neighborhood amenities has on the probability of performing PA. In order to do this, we included several variables that described availability of amenities and neighborhood characteristics that may be associated with likelihood of engaging in PA. The variables we include are: the squared miles of parks available in a 1 km radius from individuals' residential location, a first principal component index generated using variables that describe the availability of facilities within a 5 km residential buffer (public and private PA amenities, instructional, and outdoor PA-related facilities) that could support PA[3], and a crime index (number of arrests for violent crimes in the county). Some of these variables are statistically significant in the individual logit estimation, and some of them remain significant in our preferred specification (jointly estimated model). We find evidence that the availability of parks increases the probability of medium and high levels of PA for women, and the higher crime reduced the probability of high levels of PA for women as well. In the case of women there was a significant and positive effect of the index of PA related facilities on the probability of high PA levels, but just for women living in neighborhoods in the second quartile of neighborhood median income. In the case of men there was no effect of these characteristics except for the crime index, which had a positive effect in the probability of moderated PA.

5.2.1 Smoking, Childbirth, and Fast Food Meals

Table 4 presents the results of the estimation of the smoking equation, the fertility equation, and the linear model for the number of fast-food meals. As before, the table presents the equations estimated jointly with all remaining equations of the system.

Individuals are significantly more likely to smoke if they smoked in the previous period. In the case of women, the probability of smoking decreases if they practiced moderated or high levels of

[3]The variables used in the construction of the index are the following: number of public PA related amenities, number of private PA amenities that public can use by paying a fee, number of schools or academics of any PA related activity, number of outdoor places suitable for the practice of PA. All of these variables were generated for a radius of 5 kilometers with center in the individual's residential location.

Table 3A: Physical Activity Estimation Results

Variable	Women			Men			Women			Men		
	[3]: Jointly Estimated PA=2 relative to PA=1			[4]: Jointly Estimated PA=2 relative to PA=1			[7]: Jointly Estimated PA=3 relative to PA=1			[8]: Jointly Estimated PA=3 relative to PA=1		
	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx
Constant	1.39	1.09		3.66	1.36***		7.52	1.28***		6.59	1.40***	
Obese (t-1)	0.00	0.07	0.013	-0.24	0.10**	-0.047	-0.17	0.11	-0.018	-0.06	0.14	0.013
>=1 Childbirths (t-1)	0.08	0.10	-0.009				0.35	0.20*	0.033			
Smoker (t-1)	-0.04	0.07	0.018	-0.02	0.07	0.031	-0.37	0.10***	-0.037	-0.38	0.08***	-0.048
PA 3 or 4 times per week	0.56	0.07***	0.051	0.75	0.09***	0.036	0.97	0.12***	0.064	1.39	0.14***	0.111
PA 5+ times per week	0.89	0.08***	0.020	1.13	0.10***	-0.051	2.15	0.13***	0.192	2.82	0.15***	0.312
# Fast Food meals (t-1)	-0.04	0.02**	-0.003	0.01	0.02	0.003	-0.08	0.03***	-0.005	-0.01	0.02	-0.002
Family Income ($16K-$30k)	-0.09	0.08	-0.008	0.05	0.11	0.007	-0.17	0.13	-0.012	0.04	0.14	0.001
Family Income ($30K-$50k)	0.13	0.08	0.029	0.21	0.11*	0.026	0.01	0.13	-0.009	0.21	0.14	0.008
Family Income ($50K- $100K)	0.16	0.08*	0.035	0.22	0.11**	0.025	0.01	0.12	-0.011	0.26	0.13*	0.014
Family Income (+$100K)	0.22	0.10**	0.038	0.30	0.12**	0.022	0.14	0.15	-0.001	0.45	0.15***	0.033
Family Income Missing	-0.21	0.11*	-0.026	-0.28	0.13**	-0.046	-0.27	0.14*	-0.014	-0.18	0.16	0.001
High Education/ White Collars	0.04	0.10	0.007	0.00	0.15	0.032	0.01	0.17	-0.001	-0.34	0.18*	-0.045
High Education/ Blue Collars	-0.13	0.09	0.003	-0.39	0.13***	-0.026	-0.45	0.16***	-0.038	-0.69	0.15***	-0.056
Med-Low Education/White Collars	-0.21	0.09**	-0.021	-0.59	0.14***	-0.048	-0.34	0.16**	-0.022	-0.99	0.17***	-0.078
Med-Low Education/Blue Collars	-0.30	0.10***	-0.010	-0.56	0.13***	-0.040	-0.81	0.19***	-0.064	-1.04	0.15***	-0.085
Med-Low Education/Not working	-0.29	0.10***	-0.025	-0.53	0.15***	-0.043	-0.57	0.20***	-0.040	-0.90	0.18***	-0.071
Age	0.02	0.09	0.028	-0.23	0.11**	-0.003	-0.60	0.10***	-0.051	-0.51	0.11***	-0.052
Age^2	0.00	0.00	0.000	0.00	0.00**	0.000	0.01	0.00***	0.001	0.01	0.00***	0.001
African American	-0.33	0.07***	-0.028	-0.21	0.09**	-0.032	-0.63	0.10***	-0.044	-0.15	0.10	-0.001
Asian	-0.04	0.14	0.003	0.12	0.16	0.005	-0.17	0.19	-0.015	0.22	0.18	0.018
Hispanic	-0.06	0.09	0.004	-0.01	0.10	-0.006	-0.22	0.12*	-0.020	0.03	0.12	0.005
1st generation immigrant	-0.40	0.14***	-0.065	-0.29	0.16*	-0.015	-0.30	0.20	-0.004	-0.58	0.19***	-0.049
2nd generation immigrant	-0.14	0.13	-0.023	0.03	0.16	0.007	-0.11	0.16	-0.002	0.00	0.18	-0.003
Married	-0.05	0.08	0.004	-0.21	0.10**	0.007	-0.22	0.14	-0.020	-0.59	0.14***	-0.060
Cohabitating	-0.28	0.08***	-0.026	-0.17	0.09*	0.007	-0.50	0.15***	-0.034	-0.49	0.13***	-0.049
Divorced or separated	-0.16	0.08**	-0.041	-0.05	0.10	-0.027	0.07	0.15	0.020	0.19	0.13	0.030
Living with Parents	-0.10	0.08	-0.012	-0.13	0.09	-0.018	-0.13	0.13	-0.007	-0.13	0.12	-0.005
No. Children younger than 6	-0.18	0.04***	-0.014	0.01	0.06	0.012	-0.38	0.09***	-0.028	-0.11	0.09	-0.015
No. Children older than 6	0.06	0.07	0.020	0.16	0.09*	0.015	-0.10	0.16	-0.015	0.21	0.14	0.013

Table 3B: Physical Activity Estimation Results (Continued from Previous Page)

Variable	Women [3]: Jointly Estimated PA=2 relative to PA=1			Men [4]: Jointly Estimated PA=2 relative to PA=1			Women [7]: Jointly Estimated PA=3 relative to PA=1			Men [8]: Jointly Estimated PA=3 relative to PA=1		
	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx	Coeff	S. D.	Mfx
Family size	-0.01	0.02	-0.002	0.01	0.02	0.003	0.00	0.03	0.001	-0.01	0.03	-0.003
Initial H/H: Step parents	-0.17	0.07 **	-0.036	-0.06	0.09	0.001	-0.01	0.10	0.011	-0.14	0.11	-0.014
Initial H/H: Single Father	-0.11	0.16	-0.018	0.17	0.19	0.026	-0.08	0.26	-0.001	0.13	0.22	0.001
Initial H/H: Step Mother	-0.09	0.07	-0.026	0.09	0.08	0.019	0.06	0.09	0.014	0.01	0.10	-0.008
Initial H/H: Non Parents	-0.22	0.11 **	-0.044	0.29	0.17 *	0.034	-0.06	0.18	0.009	0.31	0.20	0.014
Parents Education: High School	-0.16	0.09 *	-0.017	-0.09	0.11	-0.020	-0.26	0.12 **	-0.016	0.00	0.14	0.008
Parents Education: Some College	-0.10	0.09	-0.020	-0.07	0.12	-0.021	-0.01	0.12	0.006	0.06	0.14	0.014
Parents Education: Bachelor	0.02	0.10	-0.010	-0.04	0.13	-0.010	0.18	0.14	0.019	0.00	0.15	0.004
Parents Education: +Bachelor	0.09	0.11	-0.001	0.01	0.14	0.002	0.27	0.15 *	0.023	0.01	0.17	0.000
Parents Education: Missing	-0.06	0.12	-0.001	-0.05	0.18	-0.023	-0.16	0.20	-0.013	0.13	0.21	0.022
Miles^2 of parks[1]	0.03	0.01 ***	0.005	0.01	0.02	0.003	0.02	0.01 *	0.000	0.01	0.02	-0.001
PA Related Amenities Factor Index (Pc1)	-0.03	0.05	-0.003	0.02	0.07	0.004	-0.04	0.08	-0.003	0.00	0.08	-0.001
Pc1 x {2nd Nhood Income Quartile}	0.11	0.07	0.009	-0.04	0.09	-0.010	0.19	0.08 **	0.013	0.03	0.10	0.007
Pc1 x {3rd Nhood Income Quartile}	0.04	0.06	0.002	-0.07	0.08	-0.011	0.09	0.09	0.007	-0.05	0.09	0.000
Pc1 x {4th Nhood Income Quartile}	0.01	0.06	0.007	0.00	0.08	-0.003	-0.06	0.08	-0.007	0.04	0.09	0.005
Violent Arrest by 10000 inhabitants	-0.01	0.01	0.000	0.01	0.01 *	0.002	-0.02	0.01 **	-0.002	0.01	0.01	0.000
Non PA related Amenities 5km buffers	0.04	0.07	0.011	-0.04	0.08	-0.006	-0.02	0.12	-0.006	-0.04	0.10	-0.001
Using contraception methods (t-1)	0.14	0.07 *	0.010				0.27	0.10 ***	0.020			
C2ER price of a cigarrete Carton	0.00	0.01	-0.001	0.00	0.01	-0.001	0.01	0.01	0.001	0.00	0.01	0.000
C2ER Index price for Groceries	0.16	0.15	0.018	-0.12	0.17	-0.029	0.25	0.20	0.015	0.03	0.19	0.015
C2ER Index price for Junk food	-0.11	0.09	-0.016	-0.02	0.11	-0.004	-0.11	0.12	-0.004	-0.02	0.12	0.000
Dummy third wave	-1.51	0.20 ***	-0.100	0.09	0.24	0.179	-3.24	0.26 ***	-0.254	-1.60	0.27 ***	-0.236
Dummy Fourth wave	-1.02	0.27 ***	-0.075	0.36	0.32	0.105	-2.05	0.39 ***	-0.156	-0.26	0.37	-0.070
2nd Quartile of Nhood Income	0.29	0.09 ***	0.042	-0.02	0.11	-0.002	0.29	0.12 **	0.010	-0.02	0.13	-0.001
3rd Quartile of Nhood Income	0.23	0.10 **	0.043	0.05	0.11	0.012	0.11	0.14	-0.006	-0.01	0.13	-0.006
4th Quartile of Nhood Income	0.35	0.11 ***	0.072	0.22	0.12 *	0.022	0.10	0.17	-0.016	0.28	0.15 *	0.017

Notes: *** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level. Nhood stands for Neighborhood

[1] Within 1km of Tract boundaries

Quantiles of Neigborhood income were generated using the median household income in the Census Tract. All prices are in 2005 dollars

PA. Again, this evidence suggests that unhealthy practices in the past may explain current unhealthy lifestyles. Female (male) previous smokers have an increment of 38 (47) percentage points for the probability of smoking, whereas for obese women in the last period there is an increment of 4 percentage points for the probability of smoking. Some demographic variables have a negative effect on 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 (Caucasians). First and second (only second for women) generation immigrant status have also negative effects in the probability of smoking. Married (versus single) individuals have lower probability of smoking, and there is a negative and significant effect of having children who are less than six years old for the probability of smoking for women. There is a negative effect for the probability of smoking for individuals in high income households.

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

Variable	Women				Men				Women				Women			Men		
	Smoking: Jointly Estimated [2]								Childbirth: Jointly Estimated [2]				Fast Food Meals: Jointly Estimated [2]			Fast Food Meals: Jointly Estimated [2]		
	Coef	S. D.		Mfx	Coef	S. D.		Mfx	Coef	S. D.		Mfx	Coef	S. D.		Coef	S. D.	
Constant	-1.49	1.01			-0.03	0.95			-16.34	1.70	***		12.25	0.58	***	9.62	0.52	***
Obese (t-1)	0.28	0.09	***	0.041	-0.03	0.08		-0.005	-0.42	0.09	***	-0.039	0.02	0.04		0.04	0.05	
>=1 Childbirths (t-1)	0.17	0.12		0.025					-0.25	0.11	**	-0.024	0.06	0.07				
Smoker (t-1)	2.01	0.09	***	0.379	2.20	0.06	***	0.466	0.15	0.09		0.015	-0.11	0.04	***	0.01	0.03	
PA 3 or 4 times per week	-0.18	0.08	**	-0.026	-0.07	0.08		-0.011	0.00	0.08		0.000	-0.05	0.04		0.03	0.06	
PA 5+ times per week	-0.31	0.10	***	-0.045	-0.10	0.09		-0.016	-0.14	0.10		-0.014	-0.04	0.05		0.06	0.06	
# Fast Food meals (t-1)	0.00	0.02		0.000	0.00	0.02		0.000	0.00	0.02		-0.0004	0.37	0.02	***	0.38	0.01	***
Family Income ($16K-$30k)	0.08	0.11		0.012	0.06	0.11		0.009	0.02	0.11		0.002	-0.04	0.06		0.02	0.07	
Family Income ($30K-$50k)	-0.15	0.11		-0.022	0.01	0.10		0.001	-0.03	0.11		-0.003	0.03	0.05		-0.05	0.07	
Family Income ($50K- $100K)	-0.33	0.10	***	-0.047	-0.22	0.10	**	-0.036	-0.33	0.11	***	-0.031	-0.03	0.05		0.06	0.07	
Family Income (+$100K)	-0.46	0.12	***	-0.063	-0.27	0.11	**	-0.043	-0.54	0.13	***	-0.051	-0.01	0.06		0.07	0.07	
Family Income Missing	-0.33	0.12	***	-0.046	-0.12	0.12		-0.020	-0.27	0.15	*	-0.026	-0.02	0.06		0.01	0.08	
High Education/ White Collars	-0.25	0.13	*	-0.035	-0.25	0.14	*	-0.040	-0.21	0.12	*	-0.020	-0.10	0.06	*	0.10	0.08	
High Education/ Blue Collars	0.24	0.12	**	0.036	0.46	0.12	***	0.078	0.10	0.12		0.010	0.09	0.06		0.30	0.08	***
Med-Low Education/White Collars	0.34	0.12	***	0.051	0.13	0.13		0.022	0.22	0.12	*	0.022	0.20	0.07	***	0.33	0.09	***
Med-Low Education/Blue Collars	0.56	0.13	***	0.086	0.75	0.11	***	0.132	0.38	0.13	***	0.039	0.31	0.07	***	0.44	0.08	***
Med-Low Education/Not working	0.66	0.13	***	0.104	0.85	0.14	***	0.150	1.04	0.13	***	0.113	0.00	0.07		0.16	0.10	
Age	0.11	0.07		0.017	0.02	0.06		0.004	0.99	0.13	***	0.102	0.08	0.03	**	0.29	0.03	***
Age2	0.00	0.00		0.000	0.00	0.00		0.000	-0.02	0.00	***	-0.001	0.00	0.00	**	-0.01	0.00	***
African American	-0.98	0.09	***	-0.133	-0.41	0.07	***	-0.065	0.73	0.10	***	0.074	0.28	0.04	***	0.25	0.04	***
Asian	-0.19	0.18		-0.027	-0.06	0.12		-0.009	-0.03	0.20		-0.003	0.16	0.06	***	0.13	0.06	**
Hispanic	-0.52	0.10	***	-0.071	-0.19	0.08	**	-0.031	0.17	0.12		0.017	0.18	0.04	***	0.12	0.05	***
1st generation immigrant	-1.02	0.24	***	-0.128	-0.34	0.14	**	-0.053	-0.64	0.20	***	-0.059	-0.05	0.07		-0.05	0.08	
2nd generation immigrant	-0.42	0.16	***	-0.057	-0.11	0.12		-0.017	-0.15	0.17		-0.015	-0.11	0.05	**	0.11	0.06	*
Married	-0.42	0.10	***	-0.059	-0.34	0.10	***	-0.054	1.98	0.10	***	0.234	-0.01	0.05		-0.26	0.07	***
Cohabitating	0.18	0.10	*	0.027	0.14	0.09		0.022	0.82	0.10	***	0.084	0.00	0.06		-0.16	0.07	**
Divorced or separated	0.70	0.11	***	0.110	0.44	0.10	***	0.076	1.18	0.11	***	0.123	-0.14	0.07	**	-0.17	0.08	**
Living with Parents	-0.07	0.10		-0.010	0.16	0.08	*	0.026	-1.27	0.11	***	-0.126	0.21	0.05	***	0.12	0.06	**
No. Children younger than 6	-0.24	0.06	***	-0.033	-0.02	0.06		-0.003				0.000	-0.02	0.03		0.08	0.05	

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

	Women				Men				Women				Women			Men		
									Childbirth: Jointly				Fast Food			Fast Food		
Variable	Smoking: Jointly Estimated [2]												Meals: Jointly			Meals: Jointly		
									Estimated [2]				Estimated [2]			Estimated [2]		
	Coef	S. D.		Mfx	Coef	S. D.		Mfx	Coef	S. D.		Mfx	Coef	S. D.		Coef	S. D.	
No. Children older than 6	0.07	0.02	***	0.010	0.07	0.09		0.011				0.000	-0.03	0.05		0.11	0.06	*
Family size	0.31	0.12	***	0.046	0.07	0.02	***	0.011	0.64	0.03	***	0.072	-0.01	0.01		-0.01	0.01	
Initial H/H: Step parents	0.06	0.70		0.008	0.31	0.07	***	0.053	0.21	0.09	**	0.021	-0.01	0.03		0.01	0.04	
Initial H/H: Single Father	0.19	0.11	*	0.028	0.21	0.15		0.035	0.10	0.22		0.010	-0.03	0.07		-0.18	0.07	**
Initial H/H: Step Mother	0.18	0.34		0.026	0.06	0.07		0.009	0.32	0.09	***	0.032	-0.03	0.03		0.08	0.04	*
Initial H/H: Non Parents	-0.02	0.21		-0.003	0.15	0.13		0.024	0.39	0.16	**	0.039	0.03	0.07		0.15	0.08	*
Parents Education: High School	0.00	0.22		0.000	0.11	0.10		0.018	-0.11	0.12		-0.010	0.00	0.04		-0.05	0.06	
Parents Education: Some College	-0.18	0.24		-0.026	0.12	0.10		0.020	-0.16	0.12		-0.015	-0.02	0.04		-0.04	0.06	
Parents Education: Bachelor	-0.19	0.26		-0.027	0.21	0.10	**	0.035	-0.33	0.13	**	-0.032	-0.09	0.05	*	-0.09	0.06	
Parents Education: +Bachelor	-0.09	0.36		-0.012	0.07	0.11		0.011	-0.64	0.15	***	-0.061	-0.21	0.05	***	-0.15	0.06	**
Parents Education: Missing	0.00	0.01		0.000	0.08	0.14		0.013	-0.32	0.18	*	-0.031	-0.15	0.07	**	-0.22	0.08	***
Miles^2 of parks[1]	-0.02	0.09		-0.002	-0.02	0.01	*	-0.003	-0.01	0.02		-0.001	0.00	0.01		0.00	0.01	
PA Related Amenities Factor Index (Pc1)	-0.03	0.11		-0.004	0.01	0.06		0.002	-0.22	0.10	**	-0.021	-0.01	0.03		-0.06	0.04	*
Pc1 x {2nd Nhood Median Income Quartile}	-0.11	0.11		-0.015	-0.04	0.07		-0.006	-0.04	0.11		-0.004	0.00	0.03		-0.01	0.04	
Pc1 x {3rd Nhood Median Income Quartile}	-0.08	0.10		-0.012	0.04	0.07		0.006	0.04	0.10		0.004	0.01	0.03		-0.04	0.04	
Pc1 x {4th Nhood Median Income Quartile}	-0.03	0.01	***	-0.005	0.08	0.07		0.014	0.03	0.09		0.003	0.02	0.03		-0.02	0.04	
Violent Arrest by 10000 inhabitants	0.21	0.08	***	0.031	-0.02	0.01	***	-0.004	0.01	0.01	*	0.001	0.00	0.00		0.01	0.00	*
Non PA related Amenities 5km buffers	-0.15	0.08	*	-0.022	-0.01	0.08		-0.002	-0.13	0.16		-0.013	-0.04	0.05		0.07	0.05	
Using any method of contraception (Lagged)	-0.01	0.01		-0.001					-0.42	0.09	***	-0.042	-0.04	0.04				
C2ER price of a cigarrete Carton, 2005 dollars	0.24	0.17		0.035	0.00	0.01		-0.001	0.01	0.01		0.001	0.00	0.00		-0.01	0.00	
C2ER Index price for Groceries, 2005 dollars	0.01	0.13		0.002	-0.09	0.13		-0.015	-0.28	0.20		-0.026	-0.36	0.07	***	-0.29	0.07	***
C2ER Index price for Junk food, 2005 dollars	-0.85	0.40	**	-0.116	-0.03	0.09		-0.005	-0.12	0.13		-0.012	0.01	0.04		0.08	0.05	*
Dummy third wave	-0.30	0.61		-0.042	-0.36	0.18	*	-0.057	0.87	0.29	***	0.084	-0.31	0.10	***	-0.50	0.11	***
Dummy Fourth wave					0.26	0.25		0.042	0.84	0.38	**	0.086	-0.60	0.13	***	-0.70	0.15	***
2nd Quartile of Nhood Median Income					0.01	0.10		0.002	-0.18	0.14		-0.017	0.02	0.05		0.00	0.06	
3rd Quartile of Nhood Median Income	0.10	0.18		0.014	-0.01	0.10		-0.001	-0.38	0.18	**	-0.037	0.07	0.05		-0.09	0.06	
4th Quartile of Nhood Median Income	0.06	0.20		0.009	-0.14	0.11		-0.022	-0.52	0.23	**	-0.050	0.03	0.06		-0.01	0.06	

Notes: *** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level. Nhood stands for Neighborhood

[1] Within 1km of Tract boundaries

Nhood stands for neighborhood

Quantiles of Neigborhood income were generated using the median household income in the Census Tract. All prices are in 2005 dollars

Female Add Health respondents who were obese in the previous period are less likely to have children during the current period. There is a negative effect on the probability of childbirth for women in high income households. There is a significant increase in the probability of having a childbirth for less than college educated women (versus women currently attending school), regardless of their occupation. Married and cohabiting females are more likely to have children in comparison with single women. African Americans women have a statistically significant higher probability of childbearing than white females. In addition, women with parents with college degree or higher educational attainment are significantly less likely to experience childbirth.

Female Add Health respondents consume fewer fast-food meals per week if they smoked in the previous period. Previous consumption of fast-food meals is a major factor in explaining current consumption. On average, African American, Hispanic, and Asian individuals consume more fast-food meals per week than whites. In comparison with individuals attending school, less than college educated individuals consume more fast food meals per week. Married, cohabitating and divorced men consume less fast food meals in comparison with single individuals. Respondents who still live with their parents significantly consume more fast-food meals per week, whereas individuals whose parents are college-educated consume fewer fast-food meals per week.

5.3 Simulations Using Parametric Bootstrap and Fit of the Model

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 other endogenous variables in the system that also determine obesity. Such situations make it difficult to measure the ultimate impact of any exogenous or endogenous covariates on the key outcomes. This difficulty can be overcome, however, by using simulations. In the simulations we 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)[4]. 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. Using simulation techniques we are able to get model-predicted obesity prevalence rates, in addition, we are able to perform experiments and calculate long run elasticities.

[4]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.

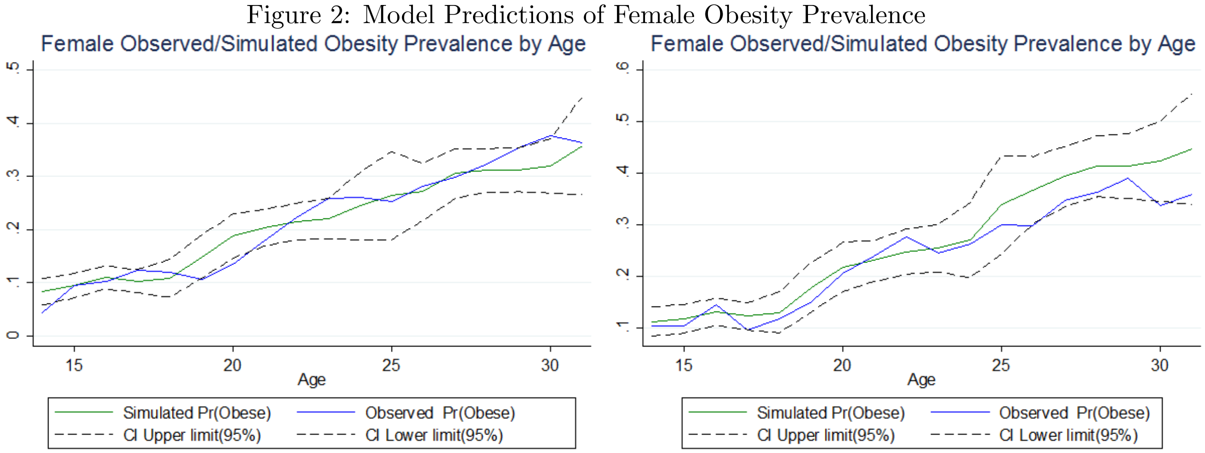
In this study we use parametric bootstrapping methods. Parametric bootstrap makes use of the following asymptotic result that holds for MLE estimation.

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

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, and Mroz, 2005). We 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.

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 Add Health study. On average, the model predicted obesity prevalence was smaller than the observed prevalence for the female Add Health respondents and greater for men. Nevertheless, the model captures the observed obesity prevalence fairly well for both women and men.

Figure 2: Model Predictions of Female Obesity Prevalence



Simulated Marginal Changes of the Obesity Prevalence Rate

Using the estimated model and the simulation techniques described in this subsection, we simulate changes in obesity prevalence as a result of changes in endogenous and exogenous variables. Five different situations are simulated using the estimated model (preferred specification). The first two relate to the individual decision to perform physical activity, the third and fourth have to do with the availability of PA-related amenities in the individual's neighborhood. The last experiment is related to the consumption of fast food meals.

In the first exercise (I) we simulate the female obesity prevalence that would have resulted if all respondents performed high levels of PA (5+ per week) from age 12-31 (simulated using the observed age-range of our data). This simulation describes an upper bound on the effect of PA on obesity prevalence for adolescents and young adults in the United States. In the second exercise (II) we simulate the obesity prevalence that would have resulted if all respondents performed high weekly PA when they were high school students. Such a situation could be the result of a policy that implements physical activity programs featuring high levels of PA in nationwide. In the third simulation (III) we increased (by one standard deviation) the availability of a set of PA related neighborhood amenities that were significant factors in the estimated equations for endogenous inputs, then using this increase in amenities, we simulate the resulting obesity prevalence. In the next exercise (IV) we increase the residential street connectivity in the participants' neighborhood in one standard deviation, by increasing the value of the beta connectivity index, and then we simulate the resulting obesity prevalence. Finally, in the last experiment we assess the effect on obesity prevalence if participants reduced their fast-food consumption by one standard deviation.

We present the results of all these exercises in Table 5, which shows the change in the obesity prevalence rate predicted in wave 4 of Add Health (when individuals were between 26 and 31 years old) and a significance test of the change. There are three columns in Table 5, each column presents a set of experiment results for a different specification. In column [1] we present results generated from our preferred specification, the jointly estimated model with residential location demand. In column 2 we present the results generated from the estimation of independent equations with no unobserved heterogeneity. In the last column in Table 5 we present experiment results generated from a jointly estimated model, but where the residential location demand is not included in the system of estimated equations. Using Table 5 we can illustrate succinctly the kind of biases that can result from assuming as exogenous the decisions that in this paper we endogenize. We first describe the results using our preferred specification and then we comment on the differences in results using different specifications.

Table 5: Experiments Based on Simulations

Experiment	[1]: Jointly Estimated Model (JEM)[3]						[2]: Estimation of Independent Equations[4]						[3]: JEM with no Residential Equation[5]					
	Women			Men			Women			Men			Women			Men		
	ΔPr(Obesity)[1]	Std. Dev	T	ΔPr(Obesity)[1]	Std. Dev	T	ΔPr(Obesity)[1]	Std. Dev	T-test[2]	ΔPr(Obesity)[1]	Std. Dev	T-test[2]	ΔPr(Obesity)[1]	Std. Dev	T-test[2]	ΔPr(Obesity)[1]	Std. Dev	T-test[2]
(I) : Intense Physical Activity throughout Adolescence and Young Adulthood.	-0.07	0.004	-16.16	-0.07	0.005	-14.56	-0.07	0.006	9.92	-0.07	0.005	-1.31	-0.07	0.004	-3.72	-0.07	0.004	1.97
(II): Intense Physical Activity in High School	-0.01	0.001	-4.49	-0.01	0.001	-4.39	-0.01	0.001	6.26	-0.01	0.001	9.02	-0.01	0.001	3.28	0.00	0.001	-6.98
(III): Increase (1 SD) in Physical Activity Related Neighborhood Amenities	-0.0021	0.001	-2.36	-0.0048	0.001	-3.93	-0.0022	0.001	2.27	-0.0023	0.001	-26.20	-0.0026	0.001	5.94	-0.0064	0.001	13.85
(IV): Increase (1 SD) in Street Conectivity Index	-0.0099	0.002	-5.95	0.0086	0.002	4.52	-0.0107	0.002	5.27	0.0043	0.002	27.90	-0.0098	0.002	-0.47	0.0067	0.002	11.64
(V) Reduction (1 SD) on the fast food meals	-0.01	0.001	-4.96	-0.03	0.003	-10.31	0.00	0.001	-83.47	-0.01	0.001	-121.18	-0.01	0.001	7.95	-0.04	0.003	43.50

Notes:

[1]This is the diference between the prevalence of obesity predited by the model under the conditions of the experiment and the prevalence of obesity predicted by the model with the observed data

[2] T statistics of a difference in means test comparing the coefficients of the spedification presented in the colunm with the coefficients of the prefered specification

[3] Experiments based on the results of our prefered specification, where all equations in the system are jointly estimated and unobserved heterogeneity is assumed to have a discrete distribution

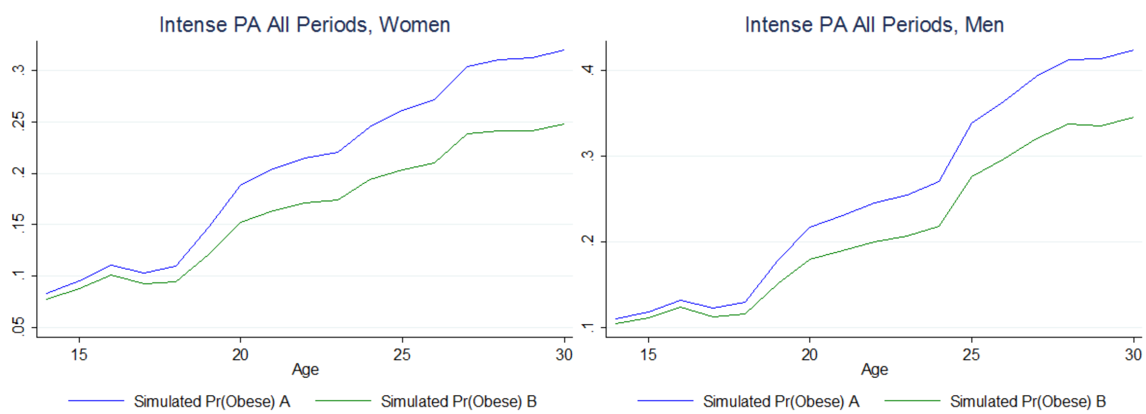
[4] Experiments based on the results of a specification where all equations in the system are independently estimated.

[5] Experiments based on the results of a specification, where all equations in the system are jointly estimated, but residential location is not modeled in any way.

This experiments are generated using a parametric bootstrap procedure with 500 repetitions.

In experiment I, we test the consequences of a generalized practice of high levels of PA during the entire observed period, in both cases (men and women), this causes a statistically significant reduction of 7% in the probability of being obese in adulthood when participants are between 26 and 31 years old (wave IV of Add Health). Figure 3 presents an illustration of the first experiment, it 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 when individuals perform high-level physical activity constantly throughout their lives (B). In experiment (II), we test the consequences of consistently engaging in high levels of PA during adolescence, when participants are in high school, finding a statistically significant reduction in adult obesity prevalence of almost one percentage point for men and women. These two exercises give us an idea of the impact of the scope of policies to encourage PA. Based on this evidence we can say that PA is an important tool that could be used to tackle the problem of obesity in the US. Reducing the consumption of fast food is another tool that can be beneficial in the reduction of obesity, from simulation number V we find that a reduction of one standard deviation in the consumption of fast food meals per week reduces the adult obesity prevalence in 0.7 and 3 percentage points for women and men, respectively. In both cases these effects are significant, and as reader may notice the reduction in obesity for men is remarkably high.

Figure 3: Experiment 1 for Women and Men



In the third simulation-based exercise we increase by one standard deviation a set of activity-related amenities, such as the square miles of parks, the first principal component index generated using variables that describe the availability of facilities that can be used to perform PA, and a one standard deviation reduction in the crime index. We find that the changes caused a reduction of 0.2 and 0.5 percentage points in the probability of adult (age 26-31) obesity for women and men, respectively. This reduction is small but it is strongly significant. In the fourth simulation exercise we increased the beta street connectivity index (a measure of degree of urban sprawl, with higher values indicating higher walkability) by one standard deviation. As a result of this change the increase in street connectivity would cause a reduction of 1 percentage point in obesity prevalence among women. In men, we find an increment of the obesity prevalence (0.8 percentage points) with higher beta street connectivity.

Almost all experiment results generated from the preferred specification (specification [1]) are statistically different from the ones obtained from specifications [2] and [3]; in most cases, the effects are overestimated using independent equations. A good illustration if this situation can be seen in experiment 2, the effect of performing high weekly PA when respondents are high school students is overestimated for men and women in specification [2] (independent equations). The same is observed in experiment 5 (a one SD reduction in fast food consumption), the effect of this experiment for women is an almost one percentage point reduction in the probability of obesity, contrary to what happens in specification [2] where the estimated effect is positive. In the case of men, specification [2] substantially overestimates the result of the fifth experiment (81 percent greater than in specification [1]). In experiment 3 (a one SD increase in the availability of a set of PA-related neighborhood amenities), we can see clearly that in comparison with the preferred specification, specification [3] overestimates the result of the experiment, the effect is 23% and 35% greater for men and women respectively. This is evidence of the importance of controlling for residential location decision, in order to estimate the effect of neighborhood amenities. In the last experiment (one SD increase in street connectivity), for women we did not find significant differences between results from specifications [1] and [3], but in the case of men the difference is significant and the model with no residential demand equation tends to underestimate the effect.

6 Conclusions

In this paper we estimate a comprehensive dynamic model of the probability of obesity that assumes endogeneity for several well recognized obesity determinants. One of the main findings is that obesity depends highly on an individual's prior weight status, which 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, our findings suggest that prior obesity increased the probability of current obesity by more than 64 percentage points. The estimated models predict that once an individual is obese, reversal is highly difficult. This is an important consideration because the prevention of initial obesity early in the lifecycle would clearly be the most efficient strategy.

This research provides evidence that one of the most important strategies for reducing obesity prevalence within a generation is through the encouragement of physical activity (PA). However, all types of PA are not equivalent determinants for reducing the probability of obesity. For women and men, we find that only high levels of physical activity (at least five times per week) significantly reduced the probability of obesity. High levels of PA physical activity were associated with a significant reduction of 4 and 6 percentage points in the probability of being obese, for women and men, respectively. Our model predicted a sizable reduction in adult obesity as a result of a continued high level PA from adolescence into adulthood (7 percentage points lower than observed). This is, in a sense, an upper bound of the potential impact that policies encouraging high levels of PA. In another exercise we simulated a situation in which all women and men performed high levels of physical activity while they were in high school, without continuation into adulthood, which resulted in a significant reduction in adult (aged 26-31) obesity by almost 1 percentage point.

In the case of women, from the set of endogenous decisions considered in the model, cigarette smoking was not a significant factor in the obesity equation estimated for women, whereas in men, smoking was associated with a reduction in the probability of obesity. Consumption of fast food meals emerged as an important obesity-promoting factor. From simulation-based experiments we found that a one standard deviation reduction in weekly fast food consumption reduced adult obesity prevalence by 0.7 and 3 percentage points for women and men respectively.

An important part of this research is testing the role that neighborhood characteristics play in encouraging PA. We did this in a framework in which the residential location decisions were explicitly modeled. Modelling the residential location is important because it helps to control for the endogeneity of neighborhood characteristics and amenities. Using the econometric framework previously described, we found evidence of a small but statistically significant effect of neighborhood amenities on reducing obesity prevalence for men and women. In simulation exercises in which we increase by one standard deviation the availability of several neighborhood amenities and reduced the crime index by one SD, the model predicted a 0.02 and 0.05 percentage point reduction in the obesity prevalence for adult men and women, respectively. In another exercise we increased street connectivity by one standard deviation, finding a significant 1 percentage point reduction in obesity prevalence for women (the effect in men was smaller in magnitude and positive). Combining results from experiments 3 and 4, in the case of women, the total effect of improving a set of neighborhood characteristics (including street connectivity) will be around 1.2 percentage point of reduction in the obesity prevalence for adult women.

Whereas some of the findings are relatively small, i.e., a reduction in obesity prevalence of 1-2 percentage points, it must be kept in mind that the mean BMI and obesity prevalence continue to rise over time in this age group , particularly among low income and socioeconomic minorities (Ogden et al. 2010). Indeed in the Add Health cohort, we found that over the 13-year study period, individuals who never developed severe obesity gained an average of 5.1 BMI units (corresponding to a 30 lb weight gain for a 5'4" woman over 13 years), whereas individuals who developed severe obesity as adults gained an average of 14.2 BMI units (corresponding to an 80 lb weight gain for a 5'4" woman over 13 years) (The, et al, 2010). Thus, the lower probability of obesity observed herein, represents a very strong impact on obesity. Furthermore, the population-impact of these changes is substantial, and could reduce the economic and health burden of obesity for millions.

Our structural dynamic model of the determinants of obesity, included well-recognized endogenous obesity determinants and accounted for the residential choice as a choice variable relevant to the young-to middle-aged adult. The whole system of equations was jointly estimated by a semi-parametric full information log-likelihood method that allowed for a general pattern of correlation in the errors across equations. A key finding is that controlling for residential self-selection has important substantive implications. Some of our simulations offer key evidence of this. For instance, in a simulation based experiment we included a one standard deviation increase in the availability of a set of PA related neighborhood amenities that in the model explains endogenous inputs, and simulated the resulting obesity prevalence. The effects on obesity prevalence from models that ignore the endogenous nature of the residential location decision overestimate the results of this experiment by 23% and 35% for women and men respectively. To our knowledge, the effects of controlling for the endogeneity of residential choice has not been yet documented within a full information maximum likelihood framework.

Acknowledgments

The authors are grateful to the Spatial Analysis Unit, and particularly Mr. Marc Peterson, at the Carolina Population Center for the bulk of the GIS data we used in this study. This work was funded by National Institutes of Health grants R01HD057194 and R01 HD041375, R01 HD39183, and a cooperative agreement with the Centers for Disease Control and Prevention (CDC SIP No. 5-00). The authors are grateful to the Carolina Population Center, University of North Carolina at Chapel Hill, for general support (grant R24 HD050924 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development [NICHD]), and to the Center for Environmental Health Sciences (CEHS), University of North Carolina (grant P30 ES010126 from the National Institute for Environmental Health Sciences [NIEHS]). NIH had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

References

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

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.

Becker, G. S. and Murphy, K. M. (1988). A theory of rational addiction. *The Journal of Political Economy*, pages 675–700.

Bhargava, A. (1991). Identification and panel data models with endogenous regressors. *The Review of Economic Studies*, 58(1):129–140.

Bhargava, A. and Sargan, J. D. (1983). Estimating dynamic random effects models from panel data covering short time periods. *Econometrica: Journal of the Econometric Society*, pages 1635–1659.

Boone-Heinonen, J. and Gordon-Larsen, P. (2011). Life stage and sex specificity in relationships between the built and socioeconomic environments and physical activity. *Journal of epidemiology and community health*, 65(10):847–852.

Chaloupka, F. J. (1991). Rational addictive behavior and cigarette smoking. Technical report, National Bureau of Economic Research.

Chou, S.-Y., Grossman, M., and 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.

Eid, J., Overman, H. G., Puga, D., and Turner, M. A. (2008). Fat city: Questioning the relationship between urban sprawl and obesity. *Journal of Urban Economics*, 63(2):385–404.

Ewing, R., Schmid, T., Killingsworth, R., Zlot, A., and 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., Carroll, M. D., Ogden, C. L., and Curtin, L. R. (2010). Prevalence and trends in obesity among us adults, 1999-2008. *JAMA: the journal of the American Medical Association*, 303(3):235–241.

French, M. T., Norton, E. C., Fang, H., and Maclean, J. C. (2010). Alcohol consumption and body weight. *Health economics*, 19(7):814–832.

Giles-Corti, B., Macintyre, S., Clarkson, J. P., Pikora, T., and 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, D. B. and Strumpf, K. S. (2005). The behavioral dynamics of youth smoking. *Journal of Human Resources*, 40(4):822–866.

Glaeser, E. L. and Kahn, M. E. (2004). Sprawl and urban growth. *Handbook of regional and urban economics*, 4:2481–2527.

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

Gunderson, E. P. and Abrams, B. (1999). Epidemiology of gestational weight gain and body weight changes after pregnancy. *Epidemiologic Reviews*, 21(2):261–275.

Harris, K. M. (2010). An integrative approach to health. *Demography*, 47(1):1–22.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320.

Keppel, K. G. and Taffel, S. M. (1993). Pregnancy-related weight gain and retention: implications of the 1990 institute of medicine guidelines. *American Journal of Public Health*, 83(8):1100–1103.

Labeaga, J. M. (1999). A double-hurdle rational addiction model with heterogeneity: estimating the demand for tobacco. *Journal of econometrics*, 93(1):49–72.

Lathey, V., Guhathakurta, S., and Aggarwal, R. M. (2009). The impact of subregional variations in urban sprawl on the prevalence of obesity and related morbidity. *Journal of Planning Education and Research*, 29(2):127–141.

Liu, H., Mroz, T. A., and Van der Klaauw, W. (2010). Maternal employment, migration, and child development. *Journal of Econometrics*, 156(1):212–228.

McFadden, D. et al. (1978). *Modelling the choice of residential location*. Institute of Transportation Studies, University of California.

Miller, W. C., Ford, C. A., Morris, M., Handcock, M. S., Schmitz, J. L., Hobbs, M. M., Cohen, M. S., Harris, K. M., and Udry, J. R. (2004). Prevalence of chlamydial and gonococcal infections among young adults in the united states. *Jama*, 291(18):2229–2236.

Mizoue, T., Ueda, R., Tokui, N., Hino, Y., and Yoshimuia, T. (1998). Body mass decrease after initial gain following smoking cessation. *International journal of epidemiology*, 27(6):984–988.

Moffit, R., Fitzgerald, J., and Gottschalk, P. (1999). Sample attrition in panel data: The role of selection on observables. *Annales d'Economie et de Statistique*, pages 129–152.

Mokdad, A. H., Marks, J. S., Stroup, D. F., and Gerberding, J. L. (2005). Correction: actual causes of death in the united states, 2000. *JAMA: the journal of the American Medical Association*, 293(3):293–294.

Mroz, T. A. (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. A. and Guilkey, D. K. (1992). *Discrete factor approximation for use in simultaneous equation models with both continuous and discrete endogenous variables.* Carolina Population Center, University of North Carolina at Chapel Hill.

Mroz, T. A. and Savage, T. H. (2006). The long-term effects of youth unemployment. *Journal of Human Resources*, 41(2):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. *JAMA: the journal of the American Medical Association*, 282(16):1523–1529.

Ng, S. W., Norton, E. C., Guilkey, D. K., and Popkin, B. M. (2012). Estimation of a dynamic model of weight. *Empirical Economics*, 42(2):413–443.

Ogden, C. L., Carroll, M. D., Curtin, L. R., Lamb, M. M., and Flegal, K. M. (2010). Prevalence of high body mass index in us children and adolescents, 2007-2008. *Jama*, 303(3):242–249.

O'Hara, P., Connett, J. E., Lee, W. W., Nides, M., Murray, R., and Wise, R. (1998). Early and late weight gain following smoking cessation in the lung health study. *American Journal of Epidemiology*, 148(9):821–830.

Parker, J., Abrams, B., et al. (1993). Differences in postpartum weight retention between black and white mothers. *Obstetrics and Gynecology*, 81(5 (Pt 1)):768.

Parsons, G. R. and Kealy, M. J. (1992). Randomly drawn opportunity sets in a random utility model of lake recreation. *Land Economics*, 68(1):93–106.

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

Rössner, S. and Öhlin, A. (1995). Pregnancy as a risk factor for obesity: lessons from the stockholm pregnancy and weight development study. *Obesity Research*, 3(S2):267s–275s.

Saelens, B. E., Sallis, J. F., Black, J. B., and Chen, D. (2003). Neighborhood-based differences in physical activity: an environment scale evaluation. *Journal Information*, 93(9).

Suchindran, C., North, K. E., Popkin, B. M., Gordon-Larsen, P., et al. (2010). Association of adolescent obesity with risk of severe obesity in adulthood. *Jama*, 304(18):2042–2047.

Wolf, A. M. and Colditz, G. A. (1998). Current estimates of the economic cost of obesity in the united states. *Obesity research*, 6(2):97–106.

Wooldridge, J. M. (2002). *Econometric Analysis Cross Section Panel*. MIT press.

Yang, Z., Gilleskie, D. B., and Norton, E. C. (2009). Health insurance, medical care, and health outcomes a model of elderly health dynamics. *Journal of Human Resources*, 44(1):47–114.

Appendix A. Unobserved Heterogeneity Parameters

Points of Support	Obesity Equation				PA=2 relative to PA=1				PA=3 relative to PA=1				Smoking Equation				Fertility		Fast-Food Equation				Pr. Weight	
	Women		Men		women		Men		Women		Men		Women		Men				Women		Men		Women	Men
	Coef	S.D	Coef	S.D	Coef	S.D	Coef	S.D	Coef	S.D	Coef	S.D	Coef	S.D	Coef	S.D	Coef	S.D	Coef	S.D	Coef	S.D		
Permanent																								
1	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.17	0.55
2	0.35	0.14	0.43	0.17	0.01	0.12	-0.08	0.12	-0.27	0.20	-0.17	0.15	-1.88	0.25	0.36	0.11	-0.42	0.16	-0.68	0.06	0.89	0.07	0.28	0.17
3	-0.21	0.31	1.47	0.29	0.10	0.14	0.03	0.16	0.15	0.24	-0.10	0.22	0.44	0.24	0.12	0.16	-0.70	0.35	-0.57	0.08	-0.07	0.04	0.19	0.28
4	0.26	0.18			-0.02	0.12			0.24	0.18			0.11	0.17			0.03	0.18	-0.63	0.07			0.36	
Time-Varying																								
1	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	0.01
2	0.40	0.51	0.66	0.50	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	-0.40	0.39	-0.41	0.32	0.39	0.38	-10.55	0.44	-11.35	0.30	0.91	0.89
3	0.28	0.44	0.43	0.42	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	-0.14	0.40	-0.20	0.33	0.34	0.39	-6.55	0.41	-7.13	0.29	0.09	0.11

Notes:

First point of support for permanent and time varying unobserved heterogeneity is normilizad to zero.

The time varying unobserved heterogeneity parameters in the physical activity equation were set to zero in order to enhance stability of the model. Models without this restriction turns out to be very unstable.

Appendix B1. Residential Location Model

Variable	Cluster 2 relative to				Cluster 3 relative to				Cluster 4 relative to			
	Women		Men		Women		Men		Women		Men	
	Coeff	S.D	Coeff	S.D	Coeff	S.D	Coeff	S.D	Coeff	S.D	Coeff	S.D
Constant	4.42	1.91	2.50	1.67	-3.49	1.89	-2.32	1.81	0.60	1.50	-0.09	1.55
Obese (t-1)	0.41	0.21	0.43	0.16	-0.38	0.13	-0.68	0.13	-0.49	0.21	-1.27	0.20
>=1 Childbirths (t-1)	0.02	0.15			-0.25	0.15			-0.43	0.22		
Smoker (t-1)	-0.36	0.13	0.12	0.08	-0.30	0.11	0.00	0.09	-0.62	0.17	-0.31	0.14
PA 3 or 4 times per week	-0.22	0.11	-0.03	0.12	0.10	0.09	0.09	0.11	0.16	0.13	0.00	0.15
PA 5+ times per week	-0.15	0.15	-0.04	0.13	-0.10	0.13	0.02	0.13	-0.05	0.19	0.12	0.19
# Fast Food meals (t-1)	0.03	0.03	0.01	0.02	-0.04	0.02	0.02	0.02	-0.12	0.03	0.03	0.03
Family Income ($16K-$30k)	-0.35	0.11	-0.36	0.11	-0.13	0.14	-0.18	0.16	-0.35	0.22	-0.51	0.23
Family Income ($30K-$50k)	-0.63	0.11	-0.58	0.11	0.09	0.13	0.10	0.14	-0.01	0.19	-0.41	0.21
Family Income ($50K- $100K)	-0.83	0.12	-0.69	0.11	0.45	0.13	0.56	0.13	0.65	0.18	0.17	0.19
Family Income (+$100K)	-1.37	0.18	-1.10	0.15	0.73	0.15	1.04	0.15	1.52	0.20	1.36	0.20
Family Income Missing	-0.46	0.16	-0.39	0.17	0.53	0.18	0.67	0.19	0.76	0.26	0.25	0.27
Age	-0.32	0.16	-0.07	0.14	-0.01	0.17	-0.07	0.17	-0.43	0.23	-0.38	0.24
Age^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00
African American	1.13	0.11	1.15	0.10	-0.45	0.11	-0.19	0.12	-0.45	0.17	0.05	0.17
Asian	-0.75	0.26	-0.20	0.21	0.19	0.19	0.41	0.18	0.58	0.25	0.98	0.24
Hispanic	0.39	0.14	0.04	0.13	0.23	0.13	0.28	0.12	0.31	0.18	0.41	0.18
1st generation immigrant	0.26	0.25	0.31	0.22	-0.07	0.22	-0.10	0.20	0.32	0.30	-0.07	0.28
2nd generation immigrant	-0.23	0.21	0.00	0.20	0.12	0.18	0.18	0.18	-0.04	0.23	0.01	0.25
Married	-0.25	0.11	-0.11	0.12	0.04	0.11	-0.02	0.11	-0.18	0.15	0.02	0.15
Cohabitating	-0.01	0.11	-0.26	0.11	-0.05	0.11	-0.34	0.11	-0.05	0.15	-0.44	0.16
Divorced or separated	-0.12	0.11	-0.15	0.11	-0.06	0.12	0.06	0.12	-0.18	0.18	0.05	0.18
Living with Parents	-0.15	0.12	-0.06	0.10	0.08	0.11	-0.01	0.11	-0.12	0.16	0.01	0.15
No. Children younger than 6	0.30	0.09	0.10	0.07	-0.14	0.07	0.04	0.07	-0.22	0.11	-0.12	0.09
No. Children older than 6	0.10	0.10	0.04	0.10	-0.05	0.10	-0.02	0.10	-0.11	0.15	0.07	0.12
Family size	-0.08	0.04	-0.07	0.03	0.05	0.03	0.00	0.03	0.05	0.04	0.02	0.04
Initial H/H: Step parents	-0.02	0.12	0.00	0.11	0.03	0.11	0.08	0.11	0.03	0.16	0.62	0.16
Initial H/H: Single Father	-0.13	0.28	0.07	0.19	0.11	0.29	-0.25	0.23	0.09	0.39	-0.34	0.36
Initial H/H: Step Mother	0.16	0.11	0.25	0.11	0.22	0.10	0.01	0.11	0.17	0.16	-0.11	0.18
Initial H/H: Non Parents	0.16	0.19	0.33	0.21	0.15	0.21	0.02	0.23	-0.36	0.32	0.00	0.36
Parents Education: High School	-0.35	0.14	-0.55	0.14	0.36	0.14	0.31	0.16	0.42	0.23	0.23	0.27
Parents Education: Some College	-0.58	0.14	-0.78	0.15	0.28	0.14	0.47	0.16	0.78	0.22	0.70	0.26
Parents Education: Bachelor	-0.59	0.15	-0.76	0.16	0.71	0.16	0.47	0.17	1.24	0.23	1.16	0.26
Parents Education: +Bachelor	-0.92	0.17	-0.49	0.17	0.55	0.17	0.77	0.18	1.22	0.25	1.40	0.27
Parents Education: Missing	-0.14	0.21	-0.51	0.21	0.25	0.23	0.28	0.24	0.77	0.34	0.64	0.36
Repeat any grade in high school	0.31	0.10	0.24	0.09	-0.43	0.11	-0.36	0.10	-0.40	0.17	-0.25	0.15
High School GPA	0.16	0.06	-0.08	0.05	0.04	0.06	-0.11	0.05	0.03	0.08	0.10	0.08
Previous contraception use	-0.01	0.10			0.07	0.11			0.23	0.16		
Wave 4 Dummy	-0.08	0.19	-0.01	0.17	1.01	0.18	0.94	0.18	1.92	0.28	2.28	0.29
Permanent UH[1]												
Parameter 1	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--
Parameter2	-0.38	0.27	-0.32	0.20	-0.58	0.20	0.33	0.19	-1.10	0.44	0.84	0.28
Parameter 3	1.24	0.27	-0.92	0.40	-1.13	0.30	1.36	0.23	-1.84	0.67	2.18	0.34
Parameter 4	-1.07	0.23			0.43	0.18			0.76	0.30		
Time-Varying UH[1]												
Parameter 1	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--	0.00	--
Parameter2	-0.19	0.52	-0.24	0.40	0.28	0.44	0.49	0.35	-0.12	0.55	0.19	0.46
Parameter 3	0.09	0.54	-0.22	0.41	0.05	0.46	0.27	0.37	-0.76	0.58	-0.38	0.49

Notes:

[1]Unobserved Heterogeneity

[1]The definition of the neighborhood-type clusters is based on a Ward cluster procedure with 4 categories

Appendix B2. Residential Model Conditional Component

Variable	Conditional Logit Coefficients (Jointly Estimated)			
	Women		Men	
	Coeff	S.D	Coeff	S.D
Population Density by Census Tract	0.0000	6.6050	0.0000	0.0000
% of African Americans by Census Tract	0.1207	4.8900	0.5255	0.1082
% any other race by Census Tract	0.1680	3.5360	0.5433	0.1590
% Hispanics by Census Tract	0.1937	-2.3410	-0.4142	0.1808
% Population<=18 by Census Tract	0.5152	-3.0040	-1.4376	0.4692
% Population>=65 by Census Tract	0.3715	-0.5810	-0.1889	0.3516
% of H/H married with children Census Tract	0.3006	1.7280	0.4794	0.2744
% Population Linguistically Isolated	0.4266	1.0040	0.3905	0.3993
% Population with less that high school	0.3175	2.3320	0.7304	0.2902
% Population with Bachelor or more	0.2042	-4.7780	-0.9310	0.1927
Median Family Income x10-4 by Census Tract	0.0000	3.8050	0.0000	0.0000
% Population receiving Assistance by Census Tract	0.6757	-2.0220	-0.9379	0.6122
2005-(median years of construction) by Census Tract	0.0016	-3.0600	-0.0052	0.0015
Median Rent by Census Tract	0.0001	-2.5970	-0.0003	0.0001
Violent Arrest by 1000000 inhabitants	0.0001	-2.0900	-0.0001	0.0001
Summation of major roads length in the Census Tract	0.0007	2.9760	0.0017	0.0006
Colleges within 3km or less of Tract boundaries	0.0134	-0.0430	0.0031	0.0126
Shopping centers within 5km or less of Tract boundaries	0.0103	-3.8770	-0.0362	0.0096
Points of interest within 5km or less of Tract boundaries	0.0053	0.3000	0.0005	0.0050
% Enrolled in College by Census Tract	0.2376	0.7060	0.1299	0.2107
% Population using public transportation	0.3001	-6.6970	-1.8630	0.2797
% Population in Urban Core	0.0556	0.4850	0.0230	0.0518
Non PA related Amenities 5km buffers	0.0049	1.6470	0.0094	0.0046
Parks within 3km Buffers	0.0067	-0.2700	-0.0021	0.0065
Principal Componet PA Amenities	0.0341	-0.6560	-0.0314	0.0319
Beta street Conectivity Index	0.2715	-1.0110	-0.3641	0.2535
ACCRA Index price for cigarretes 2005 dollars	0.0064	-1.3640	-0.0092	0.0060
ACCRA Index price for Groceries, 2005 dollars	0.1454	-0.6820	-0.0822	0.1348
ACCRA Index price for Junk food, 2005 dollars	0.0988	0.6790	0.0569	0.0912
ACCRA cost of living Index price, 2005 dollars	1.3449	-0.5910	-0.8005	1.1813

Continuing from Appendices B1 and B2

Data Appendix. Variables Descriptions and Comments

Variable	Description/Comments
Obese	Individual is obese based on the BMI criterion (BMI >30). BMI=(Weight in Kilograms)/(Height in meters)^2
At least one Childbirth in the period	The woman had at least one childbirth in the period
Smoker	Smoke at least one day during the previous month
Physical Activity 0, 1 or 2 times per week Physical Activity 3 or 4 times per week Physical Activity 5+ times per week	These variables are based on questions about how many times the individual did any of these activities last week. Bicycle, skateboard, dance, hike, hunt, do yard work. Roller blade, roller, skate, downhill ski, snow board, play racquet sports. Aerobics, strenuous team sports (such as football, soccer, basketball, etc.). Individual sports (such as running, wrestling). Gymnastics weight lifting, or strength training. Golf, fishing or bowling, softball or baseball.
Number of Fast food meal/week	How many of the past seven days individual ate food from a fast-food place
College Student	Individuals who attend to a educational institution regularly and do not have a full time job. I almost all cases their educational attainment is a college degree or more
High Education/ White Collars	Individuals with full-time/part-time white collar jobs. Their educational attainment is an associates degree, some college, or more. They do not attend to school
High Education/ Blue Collars	Individuals with full-time/part-time blue-collar jobs who are not attending school; their educational attainment is an associates degree, some college, or more
Med-Low Education/White Collars	Full-time/part-time workers with white-collar jobs who are not attending school; their educational attainment is some college or less
Med-Low Education/Blue Collars	Full-time/part-time workers with blue-collar jobs who are not attending school; their educational attainment is some college or more
Med-Low Education/Not working	Individuals with school attainment of less than high school, high school, vocational degree, or associates degree. At the time of the interview they were not working or attending school.
Age	Age
African American	The individual's race or ethnic background is African American
Asian	The individual's race or ethnic background is Asian
Hispanic	The individual's race or ethnic background is Hispanic
1st generation immigrant	Foreign born individuals with foreign born parents
2nd generation immigrant	Native born individuals with foreign born parents
Married	Individual is currently married
Cohabitating	Individual is currently cohabitating with love partner
Divorced or separated	Individual is currently divorced or separated
Living with Parents	Individual currently live with parents or family

Data Appendix (Cotinued from Previous Page)

Variable	Description/Comments
No. Children older than 6	Number of children older than 6 years old
Family size	Total members of the household
Initial H/H: Step parents	The Initial observed H/H composition comprised a stepparent
Initial H/H: Single Father	The Initial observed H/H composition comprised a single father
Initial H/H: Step Mother	The Initial observed H/H composition comprised a stepmother
Initial H/H: Non Parents	The Initial observed H/H composition comprised no parent
Parents Education: High School Parents Education: Some College Parents Education: Bachelor Parents Education +Bachelor Parents Education Missing	The education attainment of the most educated parent
Ground Transportation Terminals by County Square miles of parks within 1km of Tract boundaries Summation of major roads length in the Census Tract Colleges within 3km or less of Tract boundaries Shopping centers within 5km or less of Tract boundaries Points of interest within 5km or less of Tract boundaries	All these variables are based on GIS information from ESRI street-map products. The variables square miles of parks, shopping centers, and points of interest, were generated based on the intersection of amenity buffers with the Census Tracts polygons. The buffers are balls with center in the amenities and a predetermined radius. More information on the construction of each variable can be found in the documentation of the contextual datasets for the Add Health study (www.cpc.unc.edu/projects/Add Health). These variables are not available for wave II of the Add Health Study. In order to generate this variables for wave II, We used extrapolation methodologies based on linear functions of other neighborhood characteristics and amenities. In cases where individuals did not change location between waves II and III, variables from wave III were imputed.
Pc1: Fi…rst principal component factor index generated using variables that describe the availability of facilities within a 5 km residential buffer. The Index is based on a decomposition that generates a series of linear combinations of the variables that contain most of the	The variables used in the construction of the index are the following: instructional facilities, outdoor facilities and camps, public fee facilities that are open to the public, public facilities (public beaches, pools, tennis courts,etc.). These variables are not available for wave IV of the Add Health Study. In this case we used extrapolation methodologies based on linear functions of other neighborhood characteristics and amenities. In cases where individuals did not change location between waves III and IV, variables from wave III were imputed.
Non PA related Amenities 5km buffers	Movie theaters and arcades.

Data Appendix (Cotinued from Previous Page)

Variable	Description/Comments
Beta Street connectivity index within 5km Buffers C2ER price of a cigarettes Carton, 2005 dollars C2ER Index price for Groceries, 2005 dollars C2ER Index price for Junk food, 2005 dollars C2ER cost of living Index price	These variables are from the Obesity & Neighborhood Environment Database (ONE-Data) from the Carolina Population Center. This database provides comprehensive information on neighborhood environment measures corresponding with Add Health respondent locations in Wave I and Wave III(http://www.cpc.unc.edu/projects/onedata). More information about the generation of each variable is available in the Obesity & Neighborhood Environment Users Guide and Codebooks. In order to generate this variables for wave IV, we used extrapolation methodologies based on linear functions of other neighborhood characteristics and amenities. In cases were individuals did not change location between waves III and IV, variables from wave IV were imputed. Beta connectivity index is given by the formula: β=L/V, where L is the number of links and V the number of nodes. Higher values indicate higher connectivity.
Population Density by Census Tract % of African Americans by Census Tract % any other race by Census Tract % Hispanics by Census Tract % Population<=18 by Census Tract % Population>=65 by Census Tract % of H/H married with children Census Tract % Population Linguistically Isolated % Population with less that high school % Population with Bachelor or more Median Family Income % Population receiving Assistance Median Rent by Census Tract % Enrolled in College by Census Tract % Population using public transportation % Workers walking or bicycling to work	The construction of this variables is based on census information from Census 1990, Census 2000, and the 5-year-estimates 2005-2009 of the American Community Survey. Violent arrest were generated using the Uniform Crime Reports from the National Archives of Criminal Justice Data (www.icpsr.umich.edu). These variables were available for all waves of the Add Health study. More information on the construction of each variable can be found in the documentation of the contextual datasets for the Add Health study. (www.cpc.unc.edu/projects/Add Health).