eAPPENDIX

FOOD PRICES AND COST OF LIVING

Conducted quarterly across participating metropolitan and non-metropolitan areas (approximately 300), the survey conducted by C2ER provides price variables for more than 60 consumer goods and services including: grocery & restaurant food items, cost of living and overall price indices, and cigarette prices. Although we would have ideally included the price of alternative types of milk (i.e. low-fat milk) or away-from-home sandwiches (i.e. chicken sandwich) in addition to the price variables used, this information was not collected and thus not available to us.

To account for inflation, we used the CPI, which represents changes in prices of all goods and services purchased for consumption by urban households, including user fees and sales and excise taxes, but excluding income taxes[24]. The overall CPI is typically used in favor over food specific CPI values as it controls for total cost of living[50, 51].

Price data were linked both temporally and spatially to CARDIA respondents, as described in the main text. While respondents lived in just four U.S. metropolitan areas (representing 700 census tracts) at baseline, by year 20 they were located in 48 states, 1 federal district, 1 territory, 529 Counties and 3,805 census tracts. For individuals where there was not a direct match between residential location and city (defined as MSA) or year in which price data were collected prices were imputed. For example, if a respondent's residential location had a single matching MSA code and price data were available for the year and quarter in which the respondent was surveyed, prices from that matching MSA were assigned to the respondent.

Information on cost of living (COL) was obtained from C2ER and spatially and temporally linked to each respondent via their residential location. The COL index is based on six components – housing, utilities, grocery items, transportation, health care and miscellaneous goods and services. Data for the index were collected on more than 50,000 prices covering 60 different items, by chambers of commerce, economic development organizations or university applied economic centers in each participating urban area[23].

STATISTICAL ANALYSIS

For analysis of price elasticity (the ratio of a percent change in consumption to percent change in price), we used two step marginal effect models. These models are useful for eliminating bias when examining outcomes where there are large proportions of zero values (i.e. non-consumers) that do not represent missing data, and when the distribution of consumption is skewed[29] In the first step, the probability of consuming a particular food or beverage is estimated using a probit model with maximum likelihood estimation. In step two, a log-log ordinary least square regression model is run conditioned on consuming the food/beverage in question (i.e. uses only the subsample of consumers). Both models had the same specifications, adjusting for socioeconomic and demographic factors, logged values of other food/beverage items, logged COL, imputed price, and year. An example of the soda models are as follows:

$$\Pr(\text{AteSODA}) = \gamma_0 + \gamma_1 \text{Ln(SodaPrice)} + \gamma_2 \text{Age} + \gamma_3 \text{Female} + \gamma_4 \text{Black} +$$
$$\gamma_5 \text{LowIncome} + \gamma_6 \text{MiddleIncome} + \gamma_7 \text{Education} + \theta_8 \text{Single} +$$

$$\gamma_9 \text{MarriedWithChildren} + \gamma_{10} \text{SingleWithChildren} + \gamma_{11} \text{Ln(FoodPrices)} + \gamma_{12} \text{Ln(COL)} + \gamma_{13} \text{ImputedPrice} + \gamma_{14} \text{Year0} + \gamma_{15} \text{Year7} \quad \text{(Step 1)}$$

$$\text{Ln(SODAamt|AteSODA)} = \theta_0 + \theta_1 \text{Ln(SodaPrice)} + \theta_2 \text{Age} + \theta_3 \text{Female} + \theta_4 \text{Black} + \theta_5 \text{LowIncome} + \theta_6 \text{MiddleIncome} + \theta_7 \text{Education} + \theta_8 \text{Single} + \theta_9 \text{MarriedWithChildren} + \theta_{10} \text{SingleWithChildren} + \theta_{11} \text{Ln(FoodPrices)} + \theta_{12} \text{Ln(COL)} + \theta_{13} \text{ImputedPrice} + \theta_{14} \text{Year0} + \theta_{15} \text{Year7} \quad \text{(Step 2)}$$

To derive unconditional predicted logged consumption, the probability of consumption estimated in Step 1 is multiplied by the expected logged energy intake estimated in Step 2 in the following way:

$$E(\ln Y|X) = (\Pr(Y>0|X) \cdot E(\ln Y|\ Y>),X) = \Phi(X'\alpha) \cdot (X'\beta)$$

where Φ represents the standard normal cumulative distribution function, X' represents the vector of explanatory variables, α represents the vector of marginal effects from the probit estimation (Step 1), and β represents the vector of marginal effects from the OLS estimation (Step 2). The resulting estimates, interpreted as elasticities, are weighted means of the association between changes in price with changes in consumption for the full sample. Marginal estimates and elasticities were all generated using 1000 [bootstrapped] replications.