UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Papers in Economics

Cambridge Working Papers in Economics: 1909

NUTRITIONAL INEQUALITY: THE ROLE OF PRICES, INCOME, AND PREFERENCES

Noriko Amano-Patiño

25 January 2019

In the U.S., lower income households have a less healthy consumption basket than higher income ones. This paper studies the drivers of such nutrition inequality. I use longitudinal home-scanner data to estimate a demand system on food products, and measure the contribution of prices, disposable income and preferences to nutrition inequality. Disposable income and preferences have a predominant and quantitatively similar role in explaining consumption basket differences across income groups. Instead, prices have a limited effect. Further, I merge nutritional label information to assess, through a series of counterfactual exercises, the effect of income subsidies on nutrition quality. For example, I show that increasing the budget of a low-income household to the average level of the higher income households (a 45% increase in food expenditures) leads to an increase in protein consumption of approximately 5% and a decrease in sugar consumption of approximately 10%.

Nutritional Inequality:

The Role of Prices, Income, and Preferences.*

Noriko Amano-Patiño†

January, 2019

Abstract

In the U.S., lower income households have a less healthy consumption basket than higher income ones. This paper studies the drivers of such nutrition inequality. I use longitudinal home-scanner data to estimate a demand system on food products, and measure the contribution of prices, disposable income and preferences to nutrition inequality. Disposable income and preferences have a predominant and quantitatively similar role in explaining consumption basket differences across income groups. Instead, prices have a limited effect. Further, I merge nutritional label information to assess, through a series of counterfactual exercises, the effect of income subsidies on nutrition quality. For example, I show that increasing the budget of a low-income household to the average level of the higher income households (a 45% increase in food expenditures) leads to an increase in protein consumption of approximately 5% and a decrease in sugar consumption of approximately 10%.

*I am grateful to my advisers Joseph Altonji, Costas Meghir and Jason Abaluck for their invaluable support and to Abigail Adams, Pierre Dubois, Rachel Griffith, Aviv Nevo, Joseph Shapiro and Chen Zhen for insightful comments. I also thank Tommaso Porzio, Chiara Margaria, Rebecca McKibbin, Corina Mommaerts, Giovanni Compiani, Benjamin Friedrich, Matthew Grant, Pengpeng Xiao, Tatiana Baron, Julian Aramburu, Hugo Gerard, Marianne Bernatzky, and seminar participants at Yale for helpful comments and discussions. My research has been supported financially by the Yale Economics Department and the Yale Program in Applied Economics and Policy and by the Social Security Administration through the NBER Retirement Research Center through the "Determinants of Life Expectancy by Income & Geography: Implications for Social Security Policy" project. The analyses of this paper are based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

†University of Cambridge, noriko.amanopatino@econ.cam.ac.uk.

1 Introduction

In the U.S. there are sizable disparities in nutrition quality. For example, lower income Americans (those below 185% of the Federal Poverty Level) consume on average 20% fewer vegetables and 26% fewer whole grains per day than other Americans.[1] Poor nutrition increases the risk of cardiovascular diseases and chronic conditions and is recognized as one of the main factors contributing to the overall health disparities across income groups.[2]

Recently, due to the potential relevance for health outcomes, the stark differences in food consumption across income groups has attracted the attention of researchers. Yet, there is still little understanding of the main drivers of nutrition inequality in the U.S., which is a prerequisite for designing policies to combat the problem. This paper uses detailed household level scanner data merged with proprietary nutritional data to fill this gap. I exploit the panel dimension of household-level homescan data to separately estimate the contribution of prices, disposable income and preferences to nutritional inequality.

Using the Nielsen Homescan Data together with the Gladson Nutritional Data,[3] I first document stylized facts on nutritional disparities by income over the last decade. Second, I structurally decompose the contribution of prices, disposable income, and time-invariant heterogeneity to the differences in consumption patterns. Third, I perform counterfactual exercises to evaluate the overall effect of these components on the health quality of consumers' food baskets. Fourth, I analyze the correlation between the estimated permanent-heterogeneity parameters and demographic and socioeconomic characteristics. I outline each of these parts below, but first I briefly describe the data.

The Nielsen Homescan Data is a panel spanning the period 2004 to 2015 that contains barcode-level information of quantities purchased and prices paid – alongside household characteristics. In order to map consumption of food products into an objective characterization of nutrition quality, I merge nutritional data from Gladson's Nutrition Database and the USDA National Nutrient Database for Standard Reference. Both of these datasets record all the information contained in nutritional labels of thousands of products. I merge these databases with the Nielsen Data at the barcode level. If there is no direct match, I use the average nutritional information of products grouped by their attributes (such as brand,

[1] The disparities are even more stark for dark green vegetables, which tend to have more vitamins and minerals: lower income Americans eat 38.75% fewer dark green vegetables than higher income Americans (average intakes based on the National Health and Nutrition Examination Survey (NHANES). More details in Appendix A.1.)

[2] For example Chetty et al. (2016) document sharp differences in life expectancy by income in the U.S.

[3] The Nielsen Homescan Data are made available through an academic user agreement with the Kilts Center at Chicago Booth. The Gladson Nutritional Data was made available thanks to the Yale Program in Applied Economics and Policy.

organic claim, product description, etc.) This matching process covers approximately 97% of the products from the Nielsen Homescan Data.

Using these data I document differences in food purchases for consumption at home across different income groups. I analyze differences in consumption of healthy and unhealthy food products – as classified by Micha et al. (2017), and show that lower income households allocate a larger share of expenditure to unhealthy products with respect to higher income households. For example, a larger share of expenditure is allocated to prepared, frozen and canned foods and a lower share to fruits and vegetables as compared to higher income households. Then, I translate the differences in food consumption into a measure of nutrition quality by analyzing caloric intake and the composition of calories. I find sizable differences in both dimensions: lower income households purchase, on average, 50% more calories than the richest households. Moreover, the nutrients from which these calories are obtained differ by income. For example, the richest households source 25% more of these calories from protein and 10% less from fat and from sugar than the poorest Americans.

I then analyze consumption patterns by estimating a demand system for different food categories and an outside good; in particular, the Exact Affine Stone Index (EASI) demand system developed by Lewbel and Pendakur (2009).[4] The key source of identification to disentangle price- and income-effects from preferences – captured by time-invariant individual effects, – is cross-sectional and longitudinal variation in prices and in the level and composition of food expenditure.

The detailed level of the scanner data could allow me to define narrow food categories. However, the drawback of a fine categorization would be that each individual household would consume only a small fraction of the products leading to a censoring problem. Therefore I consider ten food categories that are detailed enough to be meaningful in terms of nutrition (according to Micha et al. (2017)) but sufficiently broad so that most households report positive consumption in each category. Namely, I focus on the categories *1)* Fruit and vegetables, *2)* Fresh meat and seafood, *3)* Unprocessed grains, *4)* Dairy, *5)* Processed meat, *6)* Processed carbs, *7)* Sweetened beverages, *8)* Frozen/Canned food, *9)* Sweeteners and desserts, *10)* Butters and Oils, *11)* Prepared food. To account for the remaining censoring in the data, I use a simulated method of moments estimator. Additionally, informed by a large literature in Industrial Organization, I address the problem of price endogeneity using average prices of goods in nearby areas to construct instruments.[5]

[4]Specifically, I use an EASI demand system that is linear in expenditure. The EASI model allows for complex dependence between the budget shares and expenditure. However, in the data, I find that adding a quadratic (or higher) order polynomial in expenditure does not improve the fit of the model.

[5]Total food expenditure may be jointly determined with expenditure shares. Following Banks et al. (1997), I use income to instrument for expenditure.

The structural estimates from the model allow me to perform a series of counterfactual exercises to characterize the impact of prices, direct income effects and preferences. In particular, I analyze the effect of equalizing the prices for all households, the effect of equalizing the preferences for all households and the effect of increasing or decreasing the households' expenditure levels while keeping preferences and prices fixed. To summarize the main results, I now describe a particular case of the latter exercise consisting of a hypothetical policy that gives subsidies for food expenditure to low-income households such that it matches the level of the highest income households (a 45% increase in food expenditures.)[6] I find that equalizing the budget for food of low-income households to the level of the richest households, a 45% increase, would make them divert their expenditure from unhealthy to healthy products, increasing the ratio of overall quantity purchased (measured in grams) of healthy to unhealthy products from approximately 60% to 67%, closing the gap with respect to higher income households by approximately 40%.[7] Specifically, an increase in food budget would decrease the consumption of "prepared, frozen and canned food" and "oil, butter and margarine" and increase consumption on the rest of the categories. Evaluating the nutritional impact of this change in consumption requires me to assess how healthy are the broad food categories I consider. I draw on the nutritional data I matched and do this using the Healthy Eating Index (a nutrition score). I find that the change in overall quantity purchased of healthy and unhealthy products translates into a two-point increase in the nutrient score of low-income households, reducing the gap with high-income households by approximately 50%. The rise in the nutrient score reflects, for example, an increase in protein consumption of approximately 5% and a decrease in sugar consumption of approximately 10%.

The other two experiments I consider consist on equalizing preferences and equalizing prices across households. Setting preferences of all households to be the average of the highest income households yields quantitatively similar effects on the nutrition score of low-income households than increasing their food budget. In contrast, a hypothetical environment in which all households face the prices of the highest income households has limited effects in the nutrient score. This is because low-income households face relatively lower prices for some of the healthy categories. These series of results suggest that policies that increase the food budget of low-income families can have a positive effect on nutrition quality.[8] At the same time, even in extreme policy cases in which the entire gap in disposable income is

[6]Notice that this exercise is feasible, in contrast with previous literature, because the framework I use allows me to analyze the pure income effect, separately from preferences.

[7]This comparison of quantities includes dairy as a healthy category as the USDA's guidelines suggest.

[8]Recent literature suggests that the marginal propensity to consume food out of food stamps (the Supplemental Nutrition Assistance Program) is higher than the propensity to consume food out of cash (Hastings and Shapiro (2017), Bruich (2014).)

closed differences in consumption remain due to differences in preferences.

In the last part of the paper I make an effort to uncover the nature of these preferences by analyzing their association with socioeconomic and demographic characteristics. Preferences for healthy products could be innately different for different income groups, or shaped by long-term exposure to characteristics that are linked to own-income or to the income level of the households' environment. I find that irrespective of the income level of their zipcode, their race and the hours of employment of the heads of the household, households with higher income have higher preferences for healthful products, suggesting that own-income plays a larger role than the environment of the household. Moreover, conditional on income, households with more educated heads of the household tend to have higher preferences for healthful foods.

1.1 Related Literature

Lower-income neighborhoods have lower availability of healthful groceries. This has led some researchers and policymakers to hypothesize that *food deserts* – that is, low-income areas without supermarkets and, hence, with fewer healthy products or more expensive ones, – are a leading cause of less healthful food consumption among lower-income households. However, Cummins, Flint and Matthews (2014), Dubowitz et al. (2015), Kozlova (2016), Allcott et al. (2018) consistently find that differences in local supply of low-income households have a limited effect on healthful consumption. This paper disentangles the portion of nutrition inequality that is unexplained by prices – the channel capturing differences in access in my setting,– into the income effect and what I refer to as the preference effect – which can be thought of as longer-term factors that affect consumption such as dietary habits.

Handbury et al. (2015) find that the causal effect of access on healthful consumption is limited. Dubowitz et al. (2015) and Cummins et al. (2014) show that the entry of a supermarket in a food desert had little impact on the consumption choices of the households in neighboring areas in the months following the opening. In concurrent work, Allcott et al. (2018) find that households are willing to travel to buy their groceries. Thus, when a new supermarket enters nearby, it may benefit households by reducing travel costs, but it does not meaningfully change their choice sets or the healthfulness of their purchases.

The literature investigating factors other than food access in explaining nutrition disparities is still sparse. Allcott et al. (2018) further embed the demand framework of Dubois, Griffith and Nevo (2014) in an equilibrium model of demand and supply and find that 91% of the nutrition-income gradient is driven by differences in demand across products, while only 9% can be attributed to differences in supply.

In contrast, I exploit the panel dimension of the data to disentangle the unexplained portion by prices into income effects and preference effects. The distinction between these two factors is relevant to develop policies aimed at equalizing nutrition quality: policies affecting the food budget of households (for example cash transfers for low income households) are likely to affect consumption through income effects. Instead, programs aimed at improving the awareness of nutrition or improving schools' lunches, could affect consumption by changing preferences. Moreover, failing to disentangle between pure income effects and preferences would lead to overestimating how much lower-income households prefer unhealthy products with respect to the healthy ones.

In the next section I describe the data sources I use. Section 3 provides evidence on nutritional in equality. In Section 4 I describe the model. Section 5 describes the estimation strategy. Estimation results and results from counterfactual exercises are reported in section Section 6. Section 7 concludes.

2 Data Sources

I base my analysis primarily on the Nielsen Homescan Consumer Panel Data made available through the Kilts Center at The University of Chicago Booth School of Business. This database is particularly well suited to analyze consumption behavior as it provides detailed purchase information (including actual prices paid and quantities bought), alongside household characteristics over the period 2004-2015.

The data is collected by a panel of households using "at-home" scanners to record all food purchases brought into the home (from department stores, grocery stores, drug stores, convenience stores and other similar retail outlets.) Participants scan each barcode and record quantity of items purchased and the store of purchase. Prices are obtained either directly from the store, if the retailer is part of Nielsen's store level data, or from the information the participant records. Thus, for each item purchased I know exactly what was bought (as denoted by the barcode or UPC), the quantity purchased, the price paid, and exactly when it was bought. The identity of the stores is not disclosed, but a 3-digit zipcode of the store, and the store chain are reported.[9]

Each participating household collects information on all products with a barcode. Items

[9]The time required to report this information raises questions about its accuracy. Einav et al. (2008) analyzed the credibility of this data and they find that, although there is non-classical measurement error in both quantities and prices, the fraction of recording errors is in line with other datasets. In particular, when comparing the Homescan data to the Nielsen Store Scanner data, they find similar recording errors to those found in earnings and employment-status from the PSID, the CEX or the CPS, for which cross-validation studies have been conducted.

without barcodes, often called "random weighted" items, are not recorded by all households. These items include fruit, vegetables, meat, and deli items. Each year there are roughly 61,000 participating households out of which a subsample of roughly 15,000 record random-weight purchases. Nielsen monitors the recording and drops households it considers are unreliable. The reliable panel, or "static" panel, has roughly 40,000 households in total, of which 8,000 a year report random-weighted purchases.

The sample I use for estimation consists of the static panel of households who record random weighted purchases. The reason for this choice is that fresh produce and fresh meat and seafood are food categories of particular interest for this project. The larger variety of products recorded comes at the cost of a smaller sample size that cannot reliably be disaggregated at the the market level as the standard Homescan sample.

Information on household demographics is collected through an annual questionnaire in which households report education, occupation and hours worked of the head, a household income range, and age and gender of all the members of the household. They also report their (five digit) zipcode, broadly defined occupations,[10] and four categories of approximate hours per week that each head is employed.[11] Further details about the Nielsen Data can be found in Appendix A.2.

The shortcoming of this data is that household characteristics are far from being thoroughly reported. In particular, nominal household incomes are reported only across discrete income ranges and those income bins are measured with a two-year lag relative to the observed shopping transactions in the dataset.[12]

The Nielsen data do not have nutritional information; I imputed this information from the data collected by Gladson. The Gladson data record information on essentially everything that is on the package of a product, including the nutritional label – which typically contains total calories, contents of saturated-, polyunsaturated- and monounsaturated-fat, total carbohydrates, sugar, fiber, vitamins, cholesterol, sodium, among many other nutrients, as well as dimensions and weight of several thousands of barcoded items. I got access to 90,000 of these items made available through the Yale Program in Applied Economics and Policy. Additionally, I drew on a similar dataset, the USDA National Nutrient Database

[10]There are twelve groups of occupations, one of them is non-employment (which includes housewife, retired, unable to work and unemployed), the rest are listed in Appendix A.11.4.

[11]The categories of approximate hours per week are "under 30 hours," "30-34 hours," "over 35 hours," and "not employed for pay." The standard number of hours for full-time employment is typically over 35 hours a week. Thus I reclassify the employment time categories into "under 34 hours," "over 35 hours" and "not employed for pay."

[12]In my estimation, I used both reported nominal incomes two years ahead to the year of the survey and contemporaneous reported nominal incomes. The results I get are very similar, so I follow Dubois et al. (2014), Handbury et al. (2015) and Jaravel (2016) and use contemporaneously reported income.

for Standard Reference, made available through the USDA. Unfortunately, this sample only covers a fraction of the products registered by the respondents of the Nielsen Homescan Data.

The linkage is further complicated by the fact that the nutrients in the Gladson Data are measured as a function of serving size. This serving size is, in many cases, measured in unconventional units (such as "one 4-inch diameter cooked pancake"), which combined with the large fraction of products that have vague number of servings per container ("varies" being the most common one) makes assessment of the matched barcodes' nutritional contents futile.

To project the nutrients available to cover a larger share of the sample, I made use the products' brand and attributes – including the description of the product (for example "canned tomato, whole") and characteristics such as organic claim, salt content and the "common consumer name description" and "variety description" (when available) in the Nielsen Homescan Data – to group the barcodes into detailed product categories that I could match with the Gladson nutritional data (based on the brand, organic claim, sodium content, and product description).[13] Details about the matching process are reported in Appendix A.2.

Finally, I use the American Community Survey (ACS) to calculate the average household income level of the Nielsen respondents at the zipcode and county level.

In the remainder of this paper, I use the terms "consumption" and "purchases" interchangeably as the Nielsen Homescan Data records purchases, which I will assume, as it is standard in the literature that uses these data,[14] equals consumption. In reality, some food may be thrown away without being consumed, or may be consumed by someone who is not a member of the household.

3 Heterogeneity in food and nutrient spending

To motivate the remainder of the paper, I now document consumption disparities across different socioeconomic groups in the U.S.

I define *nutrition quality* as the level of healthfulness of a diet. To measure this level, I follow the dietary guidelines used to construct the Healthy Eating Index (HEI)[15] and the findings of Micha et al. (2017) to classify broad groups of food products into healthy and

[13] The procedure I use to project the nutrients is similar to that followed by Dubois et al. (2014).

[14] See, for example, Dubois et al. (2014) and Handbury et al. (2015).

[15] Developed by the Department of Agriculture's Center for Nutrition Policy and Promotion (CNPP), in cooperation with Department of Agriculture's (USDA's) Food and Consumer Service and Agricultural Research Service.

unhealthy categories and to map the consumption of certain nutrients into a health score.[16]

Micha et al. (2017) found the consumption of *1)* Fruit and vegetables, *2)* Fresh meat and seafood, *3)* Unprocessed grains, to be negatively correlated with heart disease, stroke and type 2 diabetes. Instead, they found consumption of *5)* Processed meat, *6)* Processed carbs, *7)* Sweetened beverages, *8)* Frozen/Canned food, *9)* Sweeteners and desserts, *10)* Butters and Oils, *11)* Prepared food to be positively correlated with these conditions.

I use their findings and classify the first three categories as healthy and add *4)* Dairy as a healthy food category since the USDA classifies its consumption as healthy. I classify *5)* Processed meat, *6)* Processed carbs, *7)* Sweetened beverages, *8)* Frozen/Canned food, *9)* Sweeteners and desserts, *10)* Butters and Oils and *11)* Prepared food as unhealthy.

3.1 Share of health-expenditure in the cross-section

A simple comparison of the expenditure shares on the healthy categories across income groups suggests differences in consumption across income groups.[17]

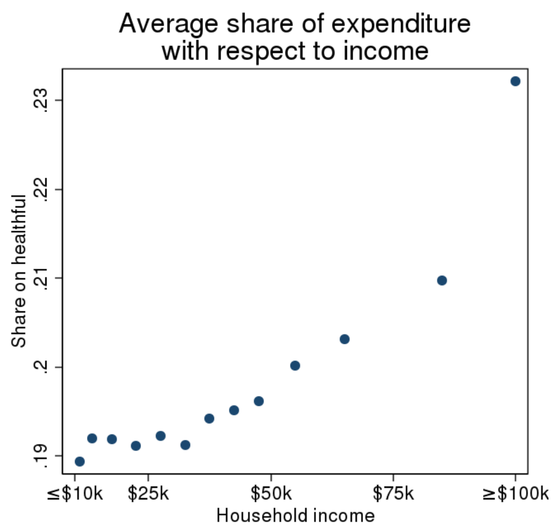


Figure 1: *Average share of expenditures allocated to healthful products.*

This comparison – showing an approximate 20% gap in expenditure in the healthy category between the top and the bottom income groups,– however, disguises many nuances

[16]This also guides the set of food products over which preferences are defined. Preferences are assumed to be separable on the seleted set of products from all other goods.

[17]More precisely, I compute the total expenditure on the healthy with respect to the unhealthy categories, and look at the share spent on each of these with respect to the total spent in these two categories as a measure of the level.

in actual food purchases. One of the many, is the prices that different income groups face: lower income households face relatively lower prices for healthy products with respect to unhealthy ones than higher income households. I analyze the relative prices for healthy products across income groups and across counties of different income levels in Appendix A.3.1. Additionally, I use quantities rather than expenditure in Section 3.3 below to show nutritional disparities across income groups.

3.2 Share of health-expenditures over time

Figure 2 plots the share of expenditure on healthy products with respect to the total spent on healthy and unhealthy by income, where each curve plots a different 2-year period. Each income group spends a different proportion on the healthy category, and although, households seem to have increased their share of expenditure on healthy goods over time in the studied period, the gap in consumption between the top and bottom income groups has remained almost constant. To capture this time trend, moving to the empirical specification I include time-fixed effects.

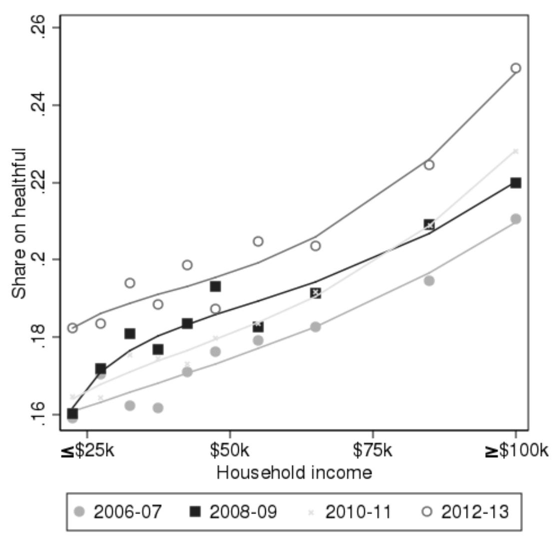


Figure 2: *Average share of expenditures allocated to healthful products.*

Next, I use nutritional information from Gladson Data to look in more detail at these disparities. I look at the total calories purchased per adult equivalent per day[18] and the share of calories obtained from saturated fats and sugars together with the intake of sodium (in milligrams per adult equivalent per day).[19]

[18]Based on the detailed adult equivalent scales by gender and age group reported by Claro et al. (2010).

[19]The reader must keep in mind that the data considered contains only food purchased for home consump-

3.3 Caloric intake

Figure 3 plots the average amount of calories purchased adjusted by the number of adult equivalents in the household per day.

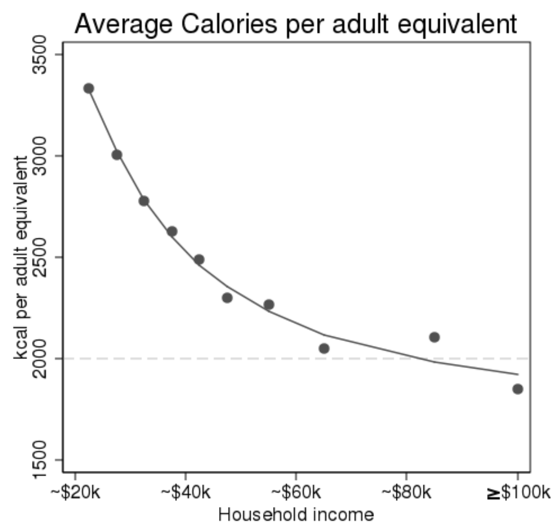


Figure 3: *Average calorie consumption per adult equivalent per day.*

Note that these are calories from at home consumption (there are no recorded food purchases in restaurants of any type). Actual caloric intake differs from what I see in the Nielsen Data. However, using a nationally representative survey that collects information about foods purchased or otherwise acquired for consumption at home and away from home, including foods acquired through food and nutrition assistance programs (FoodAPS), Handbury et al. (2015) find similar disparities across socioeconomic groups using data that incorporates food consumption outside of the home.

One would expect that households whose head spends more time outside of the household purchase more food for consumption outside of the household. However, if the patterns found by Handbury et al. (2015) using FoodAPS holds in the Nielsen Data, it should be the case that the decreasing pattern of calories with income is preserved. I verify this in Figure 4.[20]

Optimal calorie intake depends on physical activity. Therefore, I turn to analyze the composition of the calories consumed.

tion. Hence, although the mean calorie consumption per adult equivalent per day roughly coincides with the one found using dietary diaries from the NHANES (2,674 kcals per day), the actual calorie consumption is likely to be different from the one found using my data.

[20]The Nielsen Homescan Data allows for two household heads. In Figure 4, the left panel uses all observations of households in which there is only a male head, or the female head is not employed for pay. The right panel plots average calories per adult equivalent per day when there in only a female head or the male head is not employed for pay.

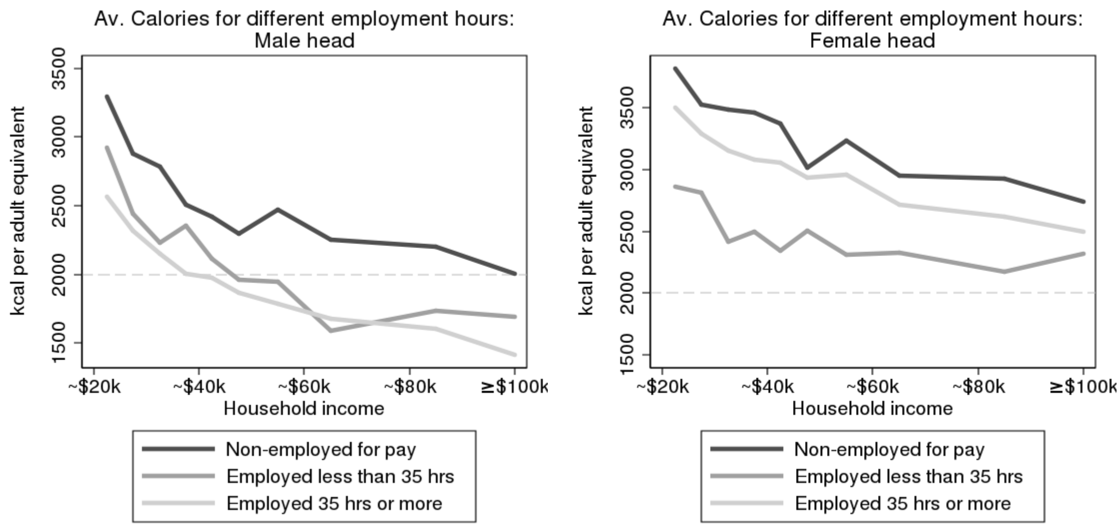


Figure 4: *Average calorie consumption per adult equivalent per day.*

The caloric content of food comprises energy in the form of the three macronutrients carbohydrates, protein and fat; total calories is a weighted sum of grams of these macronutrients.[21] Lower income households source more of their calories from fat and less of them from proteins; Figure 26 plots the average proportion of macronutrients consumed with respect to the highest income group.

Table 1: *Composition of Calories*

	Household Income	
	$\leq \$20k$	$\geq \$80k$
Calories consumed (per adult equiv/day)	3600	2000
% Calories from protein	0.112	0.118
% Calories from carbs	0.551	0.542
% Calories from sugar	0.135	0.119
% Calories from fat	0.337	0.339
% Calories Saturated fat	0.148	0.144

To summarize the differences in consumption patterns, I make use of the Healthy Eating Index: a score based on the consumption of certain products and nutrients – namely fruits and vegetables,[22] whole grains, dairy, proteins, fatty acids, refined grains, sodium and empty

[21] Approximate weights are 3.75 calories for each gram of carbohydrates, 4 calories for each gram of proteins, and 9 calories for each gram of fats.

[22] The measure used by the USDA includes fruit juice and frozen and canned vegetables as part of the "fruits and vegetables" category.

calories.[23]

Figure 5 plots the average scores by income group and confirms that the disparities we see in the expenditure shares allocated to healthful products translate into nutritional disparities as measured by the HEI. Each of the components of this index is plotted in Appendix A.4.2. The average consumption (in grams per 1000 kcal.) of "healthy" dietary components (that is, that are assigned more points for higher consumption) are increasing with income. Symmetrically, average consumption (in grams per 1000 kcal.) of "unhealthy" dietary components are mostly decreasing in income – except for consumption of processed carbohydrates.

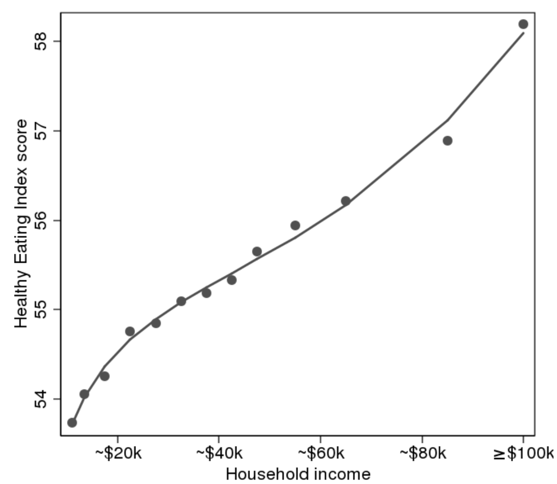


Figure 5: *Average Healthy Eating Index by income group.*

Thus far, I confirmed that low income households have, on average, a less healthful diet than higher income households using Nielsen Homescan Data.[24]

These differences in nutrient consumption could be driven by relative price differences (if lower income households face relatively lower prices for unhealthy products, they may choose less healthy goods to save money); by differences in grocery expenditures (that is, differences in the budget available to spend on food products); or by differences in preferences (where

[23]The National Institute of Health measures this component based on the proportion of calories from added sugars and saturated fats. I approximate the proportion of calories from added sugar as the proportion of calories from sugar. I report the components used in Appendix A.4.2.

[24]I used the random weighted subsample of the Nielsen Homescan Data to have more reliable information regarding purchases of fresh produce and meat. Note that the quantities reported in Table 1 are not directly comparable to those reported in Section A.1 of the Appendix as the income groups differ.

differences in preferences encompass any time-invariant component that affects demand).[25,26] To disentangle the role of each of these variables on diet quality, I estimate a demand model for 11 food categories and an outside good based on my first definition of diet quality.

4 The Model

I use a demand model to disentangle the contribution of prices, income and time-invariant idiosyncratic characteristics to explain food consumption choices. To simplify the analysis, I assume that consumers divide total expenditure, E, between expenditure on food products, x_{food}, and expenditure on the rest of the goods, x_{other}; $x_{food} + x_{other} = E$. Then, in a second stage, the expenditure of each of these broad sets of goods is allocated to the commodities within each set. I model the second stage allocation decision within the food category, thus I denote x_{food} as x. In this stage, the allocation decision is a function of only total expenditure on food products, x_{food}, and prices of products in that group. This two-stage budgeting process is equivalent to the original (one-stage) consumer problem under weak separability of preferences. That is, I assume that the utility from consumption of the different goods can be written as $f\left(v_{food}(q_1, \ldots, q_J), v_{other}(q_0)\right)$, where f is some increasing function and v_{food} and v_{other} are sub-utilities associated with food products $q_1, \ldots, q_J$ and the rest of the commodities, q_0, respectively. This approach is based on the fact that weak separability of a subset of goods from all other goods in the consumer's preference relation is necessary and sufficient for the existence of the conditional demand equations Gorman (1996).

I now explain the demand system adopted in this application. Further details about the model can be found in Appendix A.6.

The main idea of the EASI model developed by Lewbel and Pendakur (2009) is to specify a cost (expenditure) function, $c(\mathbf{P}, v_{food}, v_{other})$, that is equal to (an affine transformation of) the log-Stone Index deflated expenditures[27] for some cardinalization of the subutility function, $f\left(v_{food}(q_1, \ldots, q_J), v_{other}(q_0)\right)$, where $\mathbf{P} = (P_0, \ldots, P_J)$ denotes the vector of prices of the J food products. This approach avoids the need of recovering the indirect utility function while leaving the functional form of the Hicksian demands completely unrestricted.[28]

[25] *Differences in preferences* will capture, for example, differences in the leisure time available to prepare foods, differences in the relative prices of healthy versus of unhealthy food products, unawareness of the healthfulness of the food products they consume; time costs required to change their food basket; differences in discount rates on their future well-being, etc.

[26] I abstract from the effects of access on demand as Handbury et al. (2015), Cummins et al. (2014), Dubowitz et al. (2015), Kozlova (2016) and Allcott et al. (2017) have found limited effects.

[27] The log-Stone Index deflated expenditures are $\log(x) - \sum_{j=0}^{J} \mathrm{w}_j \log(P_j)$, where $\log(x)$ denotes total expenditures on goods, w_j the budget share allocated to product j, and $\log(P_j)$ its corresponding log price.

[28] At the cost of having the Marshallian demands implicitly defined through the cardinalization of utility.

Preferences over food products are non-homothetic because they depend on expenditures, which vary systematically by income.[29]

Let $\mathbf{w} = (w_0, \ldots, w_J)$ denote the vector of expenditure shares allocated to each of the food products and the outside good; $w_j = \frac{P_j q_j}{x}$. Let y denote log Stone Index deflated expenditures $\log(x) - \sum_{j=0}^{J} \mathrm{w}_j \log(P_j)$.

Although the model allows for complex dependence between the budget shares and expenditure, in the data I find that adding a quadratic (or higher) order polynomial does not improve the fit of the model.[30] Thus, I consider estimating equations in which expenditure enters linearly and I add time fixed-effects to account for the time trend. I obtain a system with j-th equation given by

$$\mathrm{w}_{ijt} = \alpha_j + \beta_j y_{it} + \sum_{k=0}^{J} a_{jk} \log(P_{ikt}) + \xi_{ij} + \delta_{jt} + \varepsilon_{ijt}, \tag{1}$$

where $y_{it} = \log(x_{it}) - \sum_{j=0}^{J} \mathrm{w}_{ijt} \log(P_j)$.[31] where a_{jk} captures the effect of the price of good k on the demand for good j and t indexes a half-year period and I decompose the unobservable preference component of the model $\nu_{ijt} = \xi_{ij} + \delta_{jt} + \varepsilon_{ijt}$ into a household-category specific (time-invariant) term ξ_{ij} – that captures household preferences for each of the food categories,– a time-category specific term, δ_{jt} and a random component ε_{ijt}.[32]

Real expenditure is endogenous because budget shares w_{ijt} are used to deflate total food expenditure x. Additionally, x may be jointly determined with the expenditure shares allocated to each of the products. As I discuss in Section 5, I account for this potential simultaneous equations bias by instrumenting current expenditure with current income.

[29] In Appendix A.6 I show that expenditures are monotonic in income.

[30] I tabulate the adjusted R^2's of different specifications for the polynomial on expenditure in Appendix A.6.2.

[31] Formally, under the specification in (1), real food expenditure is given by

$$y_{it} = \log(x_{it}) - \sum_{j=0}^{J} w_{ijt} \log(P_{ijt}) + \frac{1}{2} \sum_{j=0}^{J} \sum_{k=0}^{J} a_{jk} \log(P_{ijt}) \log(P_{ikt}),$$

Lewbel and Pendakur (2009) find this approximation of the deflator for nominal expenditure to provide similar estimates to those obtained using the true deflator. I opt for the use of this approximation of the EASI model as the simulated method of moments approach to deal with censoring becomes very computationally intensive.

[32] Note that in contrast with the Almost Ideal Demand System from Deaton and Muellbauer (1980), with this specification of the EASI model, household-category preferences do not enter the true deflator of nominal expenditure. This is because the underlying cost function and hicksian demands that yield the model (1) are such that the unobserved component ν_{ijt} does not enter the true deflator of nominal food expenditure, and household-specific characteristics enter only through this term.

4.1 Prices

Note that, on the one hand, prices of the food products in the data are observed at the barcode level, whereas the demand system is defined at the product category level. On the other hand, the price of the outside good is not observed. In the rest of this section I explain how I define the corresponding prices for each of the food categories and for the outside good.

A natural way to define the price indices for the food categories $\{1, \ldots, J\}$, $(P_j)_{j=1}^J$, without using the same expenditure weights for all consumers,[33] is to use a weighted average of the actual prices paid for a food product, with weights reflecting the product shares at the national level for a given year.

Let j denote a food category, i the households in the sample and t a half-year period. I define P_{ijt} to be

$$P_{ijt} = \sum_{u \in U_{jt}} \frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}} P_{it}(u),$$

with U_{jt} the set of barcodes u in category j and $\frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}}$ defined as the (nation-wide) expenditure share on barcode $u \in U_{jt}$ with respect to the rest of the barcodes in that category at time t.[34,35] For the remainder of the paper, I will refer to these price indices as *prices*.

Turning to the outside good, I define its corresponding expenditure share, w_{i0t}, as $w_{i0t} = (x_{it} - \sum_{j=1}^{J} P_{ijt} q_{ijt})/x_{ijt}$. Then, in order to calculate its corresponding price index I use quarterly prices compiled by the Council for Community and Economic Research in the ACCRA Cost of Living Index Data, for 56 goods in more than 300 U.S. urban areas.[36] Specifically, and following Zhen et al. (2013), I construct quarterly Laspeyres cost-of-living indices for the ACCRA urban areas using national average prices from 2000 as the base and item weights derived by the Council for Community and Economic Research from the 2000 Consumer Expenditure Survey.

[33] Which would have the undesirable property of implicitly assuming that preferences are homothetic within product categories.

[34] 3.06% of the prices paid by households are missing. I impute missing prices using a regression of the national average price for a barcode on indicators (and interactions) of market, quarter, characteristics of the barcode and income of the household as described in Appendix A.7.1. The predicted prices were then used to replace missing prices in the construction of price indices.

[35] The resulting price indices have a mean close to that of the average prices paid and have significant dispersion. Table 7 in Appendix A.7.2 displays the summary statistics of both actual prices paid and the price indices for some of the products in the data.

[36] A urban area corresponds to a Metropolitan or Micropolitan Statistical Area (MSA) or one or a group of counties. I refer to these urban areas as ACCRA urban areas.

I matched each household in the Nielsen data to a cost-of-living index based on their census tract. A direct match was obtained if the household lives inside an ACCRA urban area. Households outside all ACCRA urban areas were assigned to the cost-of-living index corresponding to their nearest urban area.

Let $CoLI_{it}$ denote the Laspeyres cost-of-living index for all goods and services for household i in period t. The price index for for the numèraire good was obtained by solving for P_{i0t} from $CoLI_{it} = \sum_{j=0}^{J} w_{ijt} \log(P_{ijt})$.

5 Estimation

As mention above, real expenditure, y, in equation (1) may be endogenous. Both because budget shares w_{ijt} are part of the deflator of total food expenditure x, and because the separability assumption underlying the demand model – which allows us to focus on the goods of interest as functions of prices and total expenditure on these goods,– raises questions regarding the possibility of simultaneity bias in the budget share equations of the model. Total expenditure may be jointly determined with expenditure shares on the food categories.[37] Following Banks et al. (1997), I use log-income as an instrument for expenditure.

Second, and more importantly, prices could be endogenous. This is the case whenever there are unobserved determinants of price that influence demand.[38] To account for this potential endogeneity, I use "leave out" average prices paid for each barcode[39] as instrumental variables. In particular, for each household i I define the instrument for the price he faced for barcode u, $P_{it}(u)$, to be the period average price for barcode u, excluding the average price in i's market. The identifying assumption is that prices may covary across markets due to common cost shocks but demand shocks are independent across markets.[40] Then, the

[37]Note that expenditure endogeneity issues arising from the correlation between unobserved household behavior and the allocation of food expenditure shares is somewhat alleviated by the fixed effects in the budget share equations.

[38]Notice that in this case, quantity-quality trade-off within a food category pointed out by Deaton (1988) is not a concern in my data. This is because two products with different qualities in the Nielsen Data have different barcodes. Instead, the quantity-quality trade-off may arise to extent that households who value quantity over quality are better at finding less expensive private label products than expensive name-brands.

[39]These are proxies for cost shifters in the spirit of Hausman (1996).

[40]Formally, suppose that the price that household i paid for barcode u in t is given by $P_{i,t}(u) = r_{m(i),t}(u) + e_{m(i),t}(u)$. $r_{m(i),t}(u)$ denote observed determinants of price that may depend on the market $m(i)$, and $e_{m(i),t}(u)$ denote unobserved determinants. The identifying assumption is that household specific demand shocks are independent across markets so that $e_{m(i),t}(u) = \eta_{m(i)}(u) + \epsilon_{m(i),t}(u)$ where the shocks $\eta_{m(i)}(u)$ are independent across markets.

"leave-out" average price paid for barcode u,

$$\pi_{i,t}(u) = \frac{1}{|M-1|} \sum_{m \neq m(i)} \frac{1}{|m|} \sum_{\ell \in m} P_{\ell,t}(u)$$

is a valid instrument for $P_{i,t}(u)$. The instrument for the price index $P_{ijt} = \sum_{u \in U_{jt}} \frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}} P_{i,t}(u)$, is defined as a weighted average of the price instruments corresponding to the barcodes that were purchased by households ℓ in i's county $c(i)$, $\Pi_{c(i),jt} = \frac{1}{|c(i)|} \sum_{\ell \in c(i)} \sum_{u \in U_{jt}} \frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}} \pi_{\ell,t}(u)$.

Let I_{it} denote i's real income in period t. Following Banks et al. (1997), I use $\log(I_{it})$ to instrument for expenditure.That is, I impose the identifying assumption that $\mathbb{E}(\boldsymbol{\varepsilon}_{it} \mid I_{i1}, \ldots, I_{iT}; \boldsymbol{\Pi}_{c(i),1}, \ldots, \boldsymbol{\Pi}_{c(i),T}) = 0$.[41]

The model is then given by the system of J structural equations in (1) together with the first stage regressions

$$\begin{aligned} \mathbf{p}_{it} &= \mathbf{b}^p \widetilde{I}_{it} + \mathbf{C} \log(\boldsymbol{\Pi}_{c(i),t}) + \mathbf{e}^p_{it} \\ y_{it} &= \mathbf{b}^y \widetilde{I}_{it} + \mathbf{F} \log(\boldsymbol{\Pi}_{c(i),t}) + \mathbf{e}^y_{it}. \end{aligned} \tag{2}$$

5.1 Dealing with Censoring

A significant fraction of households do not consume some of the broad food categories used for the analysis, which restricts the sample that can be used to estimate the model by almost 50% in each period. The literature offers several utility consistent approaches to deal with censoring. In a setting where every food product can be censored, however, these approaches are computationally infeasible. Instead, to account for corner solutions, I use a simulated method of moments estimator developed by Gourieroux et al. (1993).

The Nielsen Homescan data is an unbalanced panel in which respondents' mean duration in the sample is of 6 years. The minimum length of time they stay in the sample is 1 year. The maximum 11 – which is the total length of the panel. Most households consume at least one of the products I consider for the analysis in each period in which they belong to the sample. I restrict the sample to household-period pairs (i, t) (where a period spans half a year) in which at least one of the products from each of the groups that I consider is consumed. Only 53% of these household-period observations correspond to a bundle of the J goods in which all of the expenditure shares are positive. Let the set of observations

[41] Note that the term $\sum_{j=1}^{J} \boldsymbol{\pi}_{c(i),j,t} \bar{\mathrm{w}}_{jt}$ is not a linear transformation of $\log(\boldsymbol{\pi}_{c(i),j,t})$. The validity of this instrument lies in the assumption that households face variation in income over time, while the rest of their characteristics – in particular their consumption environment, – remains unchanged. In Section A.10.3, I use different subsamples (of the positive sample) to estimate the parameters as a robustness check.

with this property be the *positive sample*. Details about the positive sample can be found in Appendix A.8.1.

Lee and Pitt (1986) characterize household consumption behavior under censoring building on Neary and Roberts (1980).[42] Their approach uses the individual's Kuhn–Tucker conditions to construct *virtual prices*, which are reservation prices that would induce households to consume exactly zero without imposing the non-negativity constraints. They compare the virtual prices implied by corner solutions to the actual prices to derive the likelihood of observing the households' actual chosen bundle of goods.[43] This approach becomes computationally infeasible as the number of censored goods grows.

A Tobit estimator (which would circumvent the computational difficulties of the approach developed by Lee and Pitt (1986)) for the case of corner solutions in demand system estimation, results in biased estimates (for systems with more than two goods). The problem of this approach is that it fails to consider that consumers response to price depends on the set of goods it consumes at corners. To deal with the significant presence of zero expenditure on some goods without resorting to the computationally intensive approach developed by Lee and Pitt (1986), I use a simulated method of moments estimator (Gourieroux et al. (1993)).

5.1.1 Indirect Inference

The model in (1) defined over the whole sample of households requires an additional assumption on the distribution of the preference parameters ξ_{ij} that specifies the dependence of these parameters on the rest of the covariates. I specify the full *structural model* to be given by,

$$\begin{aligned} \mathbf{w}_{it} &= \boldsymbol{\beta} y_{it} + \mathbf{A} \log \mathbf{p_{it}} + \boldsymbol{\xi}_i + \boldsymbol{\delta}_t + \boldsymbol{\varepsilon}_{it}, ^{44} \\ \boldsymbol{\xi}_i &\sim N(\boldsymbol{\zeta}_0 + \boldsymbol{\zeta}_1 \bar{y}_i, \Sigma_\xi). \end{aligned} \tag{3}$$

and denote the vector of all the parameters with $\boldsymbol{\gamma}$.[45]

Let N denote the number of individuals and T the length of the panel in the observed data given by observations of expenditure shares, real income and prices paid $\{w_{ijt}, y_{it}, p_{ijt}, d_{jt}\}$,

[42] Neary and Roberts (1980) characterized more generally household consumption behavior under quantity constraints.

[43] These virtual prices are themselves functions of the household's income and of the prices of the uncensored goods. They can be calculated from the unconstrained demand and supply functions as I illustrate in Appendix A.8.3. Further details about the approach developed by Lee and Pitt (1986) can be found in Appendix A.8.2.

[44] Together with the first stage equations in (2).

[45] Specifically $\boldsymbol{\gamma} = (\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1, \overline{\Sigma}_\xi, \boldsymbol{\beta}_r, \mathbf{A}, \boldsymbol{\delta}_t)$.

$i = 1, \ldots, N$; $t = 1, \ldots, T$; $j = 1, \ldots, J$, where w_{ijt} is endogenous to the model and the fitted values $(\hat{p}_{ijt}, \hat{y}_{it})_{i,j,t}$, from the first stage equations $\mathbf{p}_{it} = \mathbf{b}y_{it} + \mathbf{C}\log(\boldsymbol{\pi}_{it}) + \mathbf{e}_{it}$ and $y_{it} = \mathbf{b}^y\tilde{y}_{it} + \mathbf{F}\log(\boldsymbol{\pi}_{it}) + \mathbf{e}^y_{it}$ are exogenous.

The *auxiliary model* is defined over the positive sample[46] and consists of the within transformation of the demand system in (3). Let the vector of parameters from the auxiliary model be denoted by $\boldsymbol{\theta}$. Note that $\boldsymbol{\theta} = \left(\boldsymbol{\zeta}^+_0, \boldsymbol{\zeta}^+_1, \overline{\Sigma}^+_\xi, \boldsymbol{\beta}^+_r, \mathbf{A}^+, \boldsymbol{\delta}^+_t\right)$ can be estimated using the observed data. In particular, I use the within transformation and estimate the auxiliary model using two stage least squares. Let $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\zeta}}^+_0, \hat{\boldsymbol{\zeta}}^+_1, \hat{\overline{\Sigma}}^+_\xi, \hat{\boldsymbol{\beta}}^+_r, \hat{\mathbf{A}}^+, \hat{\boldsymbol{\delta}}^+_t\right)$.[47]

Given $\{\hat{y}_{it}, \hat{p}_{ijt}, , d_{jt}\}_{i=1,\ldots,N;t=1,\ldots,T;j=1,\ldots,J}$ and assumed values of $\boldsymbol{\gamma}$, I use the structural model to generate M statistically independent simulated data sets $\{\widetilde{w}^m_{ijt}(\boldsymbol{\gamma}), \hat{y}_{it}, \hat{p}_{ijt}, , d_{jt}\}$, $m = 1, \ldots, M$. Each of the M simulated data sets has N individuals and is constructed using the same observations on the exogenous variables $\{\hat{y}_{it}, \hat{p}_{ijt}\}_{i,j,t}$. For each of the M simulated data sets, I estimate $\tilde{\theta}_m(\boldsymbol{\gamma}) = \left(\hat{\boldsymbol{\zeta}}^m_0, \hat{\boldsymbol{\zeta}}^m_1, \hat{\Sigma}^m_\xi, \hat{\boldsymbol{\beta}}^m_r, \hat{\mathbf{A}}^m, \hat{\mathbf{d}}^m_t\right)$ exactly as I estimate $\hat{\boldsymbol{\theta}}$: using the within transformation of the model and two stage least squares. Denote the average of the estimated parameter vectors by $\tilde{\theta}(\boldsymbol{\gamma}) = \frac{1}{M}\sum_{m=1}^M \tilde{\theta}_m(\boldsymbol{\gamma})$. Indirect inference generates an estimate $\hat{\boldsymbol{\gamma}}$ of the structural parameters by choosing $\boldsymbol{\gamma}$ to minimize the distance between $\hat{\boldsymbol{\theta}}$ and $\tilde{\theta}(\boldsymbol{\gamma})$ according to some metric. I take the Wald metric and follow Altonji et al. (2013), adding the proportion of zeros together with the average level of expenditure shares in each of the food categories as additional moments to match,

$$
\begin{aligned}
\hat{\boldsymbol{\gamma}} = \arg\min_{\gamma} \Big[& \left(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}(\boldsymbol{\gamma})\right)' \Omega \left(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}(\boldsymbol{\gamma})\right) \\
& + \omega_{zeros}(Zeros^{\text{data}} - Zeros(\boldsymbol{\gamma})^{\text{sim}})^2 + \omega_{\overline{\mathbf{w}}}(\overline{\mathbf{w}}^{\text{data}} - \overline{\mathbf{w}}(\boldsymbol{\gamma})^{\text{sim}})^2\Big].
\end{aligned} \tag{4}
$$

where ω_{zeros} and $\omega_{\overline{\mathbf{w}}}$ denote scalar weights. Gourieroux et al. (1993) showed that the resulting estimate $\hat{\boldsymbol{\gamma}}$ is a consistent and asymptotically normal estimate of the true parameter $\boldsymbol{\gamma}$.

The vector of structural parameters $\boldsymbol{\gamma} = \left(\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1, \overline{\Sigma}_\xi, \boldsymbol{\beta}_r, \mathbf{A}, \mathbf{d}_t\right)$ has dimension $J(3 + R + J + (T-1))$. I impose symmetry of the Slutsky matrix to reduce number of price effects. This reduces the dimension of $\boldsymbol{\gamma}$ to $J(3+R+(J+1)/2+(T-1))$. Adding the fraction of zeros and the average levels of expenditure shares, yields a highly non-convex objective function with many local optima making the minimization in (9) difficult to compute. Thus, I use a simulated method of moments based on the MCMC algorithm developed by Chernozhukov and Hong (2003). This derivative-free procedure is computationally attractive because it

[46] I refer to the sub-sample of household-semester combinations for which consumption of all goods is positive as the *positive sample*.

[47] For short panels such as the one at hand, $\hat{\boldsymbol{\xi}}^+$ suffers from sampling bias.

can deal with non-smooth objective functions and a high-dimensional parameter vector.[48]

I find the bias due to censoring to be small under the fixed effect specification I use. In a cross-sectional analysis, the bias due to censoring could be sizable as it is suggested from the discussion in Appendix A.8.4.

6 Results

In this section I report the elasticities implied by the estimates of the structural parameters and perform a series of counterfactual exercises to characterize the impact of prices, direct income effects and preferences. I then analyze the correlations of preferences with socioeconomic and demographic characteristics.

6.1 Estimation Results

The *budget elasticities of demand*, that is, the elasticities of demand with respect to disposable income or the expenditure level, are given by $\epsilon_j^y = 1 + \beta_j/\mathrm{w}_j$. The first column of Table 2 reports the implied budget elasticities of demand.

Food categories have an income elasticity below 1 and the income elasticity for the outside good is in line with what is found in the literature (see Zhen et al. (2013)). Note that the table reports median income elasticities; however, income elasticities vary by income level (as expenditure shares w_j vary with income level). Interestingly, dividing the sweetened beverages category into "juice" and the rest of the beverages (that is mainly composed by soft drinks), we see a very different response to income. In particular, juice has a median income elasticity of 0.3405 and the rest of the beverages of 0.1424.

These estimates imply that, upon an increase in disposable income, consumption of the food categories would increase to different degrees. The food categories with the highest income elasticities are produce, processed carbs, dairy and unprocessed grains. The rest of the categories respond by less to changes in disposable income. In order to quantify the overall effect of a change in disposable income on the quality of consumption, I make use of the Healthy Eating Index in Section 6.2.2.

The Hicksian (compensated) price elasticities are given by $\eta_{jk} = \frac{a_{jk}}{\mathrm{w}_j} + \mathrm{w}_k - \mathbb{1}(j = k)$, and are reported in the second column of Table 2. All own compensated price elasticities are negative, however, the eigen-value corresponding to the first category, "processed meat,"

[48]Details about the MCMC algorithm can be found in Appendix A.9.

Table 2: Median Income Elasticities.

	Income Elasticities	**Own Price Elasticities**	
		Hicksian	Marshallian
	$\epsilon^y_j = 1 + \frac{\beta_j}{\bar{\mathrm{w}}_j}$	$\eta_{jj} = \frac{a_{jj}}{\bar{\mathrm{w}}_j} + \bar{\mathrm{w}}_j - 1$	$\eta_{jj} - \bar{\mathrm{w}}_j \epsilon^y_j$
Outside good	1.0741	-0.0086	-0.9667
Processed meat	0.2783	-0.0074	-0.0096
Fresh meat and seafood	0.1519	-0.0025	-0.0033
Unprocessed grains	0.3393	-0.6739	-0.6753
Processed carbs	0.3821	-0.4966	-0.5050
Sweetened beverages	0.2417	-0.9491	-0.9514
Produce	0.3987	-0.4298	-0.4365
Dairy	0.3587	-0.7972	-0.8042
Frozen/Canned food	0.2262	-0.4856	-0.4859
Sweeteners, Sweets	0.3072	-0.0128	-0.0149
Butter/Margarine/Oils	0.0946	-0.2979	-0.2980
Prepared food	0.3198	-0.9982	-1.0020

Note: This table reports the median income elasticities of demand and Hicksian and Marshallian own-price elasticities calculated using the estimates of the structural parameters for each category $j \in \{1, \ldots, J\}$. The point estimates are reported in Table 10 in Appendix A.10.2.

is positive for all income groups. Thus, the Slutsky matrix is not negative semidefinite. The third column, reports the implied Marshallian price elasticities given by $\eta_{jk} - \mathrm{w}_j \epsilon^y_k$.

As discussed in Section 5, following Banks et al. (1997), I use log-income as an instrument for expenditure. The validity of this instrument, lies in the assumption that households face variation in income over time, while the rest of their characteristics – in particular their consumption environment, – remains unchanged. As a check of the validity of this assumption, in Appendix A.10.3, I use different subsamples (of the positive sample) to compare the estimates of the parameters.

6.2 Interpretation of the results

With these estimates of the structural parameters, I evaluate alternative environments and policies to understand how they may affect consumption behavior.

The predicted expenditure shares by households in income group I are given by

$$\mathbb{E}(\mathrm{w}_{ijt} \mid I_{it} = I) = \hat{\beta}_j \mathbb{E}(y_{it} \mid I_{it} = I) + \sum_{k=1}^{J} \hat{a}_{jk} \mathbb{E}(\log(p_{ikt}) \mid I_{it} = I) + \mathbb{E}(\hat{\xi}_{ij} \mid I_{it} = I) + \hat{\delta}_{tj}.$$

I denote $x(I) = \mathbb{E}(x \mid I_{it} = I)$ and $p_{ijt} = \log(P_{ijt})$ in the remainder of the paper for ease of notation. Note that although the shares w_{ijt} are linear functions of real expenditure, they are not necessarily linear functions of income (since real expenditure on food is a nonlinear function of income).[49]

6.2.1 Price Effects

I start by analyzing the effect of equalizing relative prices across households. I set the relative prices equal to the average relative prices paid by the highest income households, $(p_j^H)_{j=1}^J$, for all households. The counterfactual expenditure shares at this price level is given by $\mathbb{E}(\mathrm{w}_{ijt} \mid I_{it} = I, p_{ikt} = p_k^H) = \hat{\beta}_j y_{it}(I) + \sum_{k=1}^J \hat{a}_{jk} p_k^H + \hat{\xi}_{ij}(I) + \gamma_{jt}$. At each income level I, the difference

$$\mathbb{E}(\mathrm{w}_{ijt} \mid I_{it} = I, p_{ikt} = p_k^H) - \mathrm{w}_{ijt}(I) = \sum_{k=1}^{J} \hat{a}_{jk} \left(p_k^H - p_{ikt}(I) \right). \tag{5}$$

captures the adjustment in expenditure share allocated to category j if group I faced prices $p_1^H, p_2^H, \ldots, p_J^H$. That is, the right-hand side of (10) captures the average effect on i's consumption of category j of "changing" the relative price schedule from $(p_{ijt}(I))_{j=1}^J$ to the average relative price schedule from the top income group $(p_j^H)_{j=1}^J$.

I then recover counterfactual quantities from these counterfactual shares by fixing expenditure on food products, $\sum_{j=1}^{11} P_j q_j$, at a hundred dollars for all households.

I summarize the overall differences using the classification into healthy and unhealthy products from Section 3 and plot the ratio of quantities bought of healthy products with respect to quantities bought of unhealthy products in Figure 6. Note that these ratios are not directly comparable to the ones in Section 3: these ratios compare quantities in grams rather than expenditure shares. The average of these differences for each income group and each product can be found in Appendix A.10.4.

As mentioned in Section 3.1 (and discussed in Appendix A.3), higher income households face relatively higher prices for healthy products. Hence, increasing the relative prices of healthy products to the level of higher income households, does not induce them to significantly divert expenditure from unhealthy to healthy products.

To measure the overall effect on nutrition of this change in food baskets' composition I use the Healthy Eating Index (HEI), which I construct using the nutrients obtained from the predicted quantities consumed of each good $q_1, q_2, \ldots, q_{11}$. The predicted nutrients are computed under the assumption that households do not change the nutritional composition of the set of barcodes they were buying within each category. That is, I assume that upon an

[49] Appendix A.10.1 shows the relationship between these two variables.

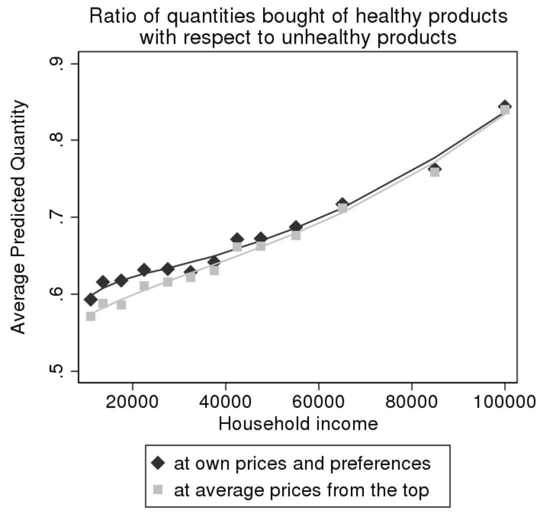


Figure 6: *Predicted ratio of overall quantities purchased (measured in grams) of healthy to unhealthy products by income group at own prices and at average prices from the highest income group. The comparison illustrates the effect of equalizing prices at the average level of the highest income group on the composition food baskets. Appendix A.10.4 shows the effect product by product.*

increase in consumption for a category, nutrients increase proportionately. For example, it may increase the quantity of dairy it consumes but does not switch from buying low fat milk to buying whole milk.[50] I thus use the average nutritional content for each food category and for each household, to impute average nutrients to the food basket $\{q_1, q_2, \ldots, q_{11}\}$. I adjust these quantities proportionally so that the total calories from the proportional basket $\{\widetilde{q}_1, \widetilde{q}_2, \ldots, \widetilde{q}_{11}\}$ contains exactly 1000 calories and I compute the HEI.The percentage difference in predicted HEI scores between the top and bottom income groups is of 8.2%.

I compare the indices obtained by constructing these counterfactual baskets using the predicted shares when prices are held at the average level of the highest income group, $\mathbb{E}(\mathrm{w}_{ijt} \mid I_{it}, p_{ikt} = p_k^H)$. The average of these indices by income group is plotted in Figure 7.

The difference in HEI scores between the top and bottom income groups reduces only slightly under the counterfactual price schedule by 8.03%. These small differences are driven by the higher relative prices for some healthy products faced by higher income households: the counterfactual prices effectively raise the relative prices for some healthy products for lower income households – in particular, fruit and vegetables.

This counterfactual exercise must be interpreted with caution, however. Notice that I use aggregated groups of products and compute the prices of these aggregated groups by using weighted averages of the prices of the products in each group. Hence it could be the

[50] Or, if it does, then it lowers the amount of fat from other products in the dairy category in such a way that the proportion of fat consumed from the dairy category remains constant.

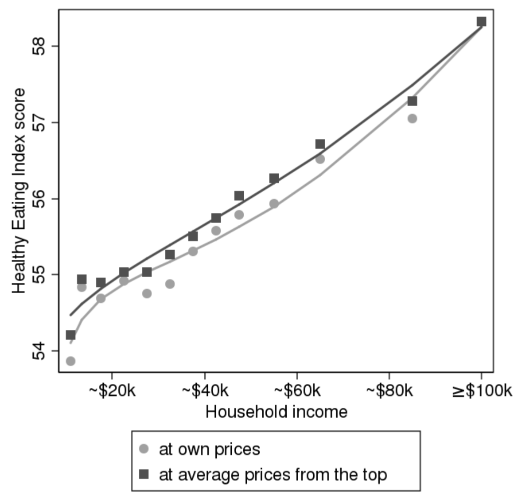


Figure 7: *Average predicted Healthy Eating Index by income group at own prices and at average prices from the highest income group. The comparison illustrates the effect on the nutrient scores of equalizing prices at the average level of the highest income group.*

case that the higher relative price for the healthy category is a result of the way I aggregate products rather than a true difference in relative prices. Indeed in Appendix A.3.2 I find evidence suggesting that higher income households buy different and more expensive goods than lower income households; more so in in the healthy category. Higher income households could face relatively higher prices for healthy products if the higher demand for healthier products in the area they lived led to an increase in prices. I compare relative price indeces of the healthy and unhealthy categories by average household income level in Appendix A.3.1. At the county level, there does not seem to be a strong correlation between average household income level and the relative prices of healthy versus unhealthy products.

6.2.2 Budget Effect

The second counterfactual exercise I do is to implement a hypothetical policy that gives subsidies for food expenditure to low-income households. To do this, I fix the prices and preferences at the average level of the lowest income households and look at the predicted expenditure shares at different expenditure levels; in particular, the average expenditure levels of different income groups.

Figure 8 plots the ratio of amounts purchased of healthy with respect to unhealthy products.

There are two features worth noticing in Figure 8. First, the joint effect of prices and preferences is captured as the vertical distance between the light gray curve and dark gray

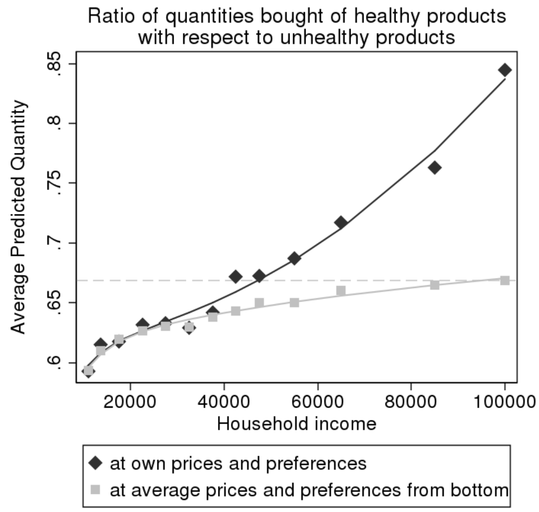


Figure 8: *Predicted ratio of overall quantities purchased (measured in grams) of healthy to unhealthy products by at prices and preferences from the lowest income group. The comparison with the predicted ratios by income groups illustrates the effect of increasing the food budget of the lowest income households to the average food expenditure level of different income groups on the composition food baskets. Appendix A.11.3 shows the effect product by product.*

curve. This effect leads to an increase in the relative quantity of healthy to unhealthy products except for the highest income households. Second, the distance between the light gray curve and the dashed line, captures the income effect on the aggregated healthy category. Let $R(y_{it}, P_{ijt}, \xi_{ijt})$ denote the relative quantity purchased of goods in the healthy to unhealthy categories for household i. The averages $\frac{1}{|I|}\sum_{i\in I} R(y_{it}, P_{ijt}, \xi_{ijt})$, which I denote for simplicity by $R(y^I, P^I, \xi^I)$, for each income group are scattered in dark gray in Figure 8. I denote with L the lowest income group, that is the group of households with incomes under \$20,000. Similarly, I use H to denote the highest income group, which is the group of households with incomes above \$100,000. The counterfactual averages $R(y^I, P^L, \xi^L)$ are plotted in the light gray curve. The distance between $R(y^L, P^L, \xi^L)$ and the dashed line, through $R(y^H, P^L, \xi^L)$, captures the income effect from shifting disposable income from the one of low-income level households, L, to the level of high-income level households, H.

The proportion of of the budget allocated to food products with respect to that corresponding to the outside good depends on the income elasticity each income group has for the outside good. However, for all income groups it implies an increase in food expenditure; more pronounced for lower income groups and less pronounced for higher income groups. $R(y^I, P^L, \xi^L) \leq R(y^{I'}, P^L, \xi^L)$ for $I < I'$ implies that the healthy category is a normal good relative to disposable income for food – rather than to income.

Indeed, consumers in this model can be thought of as allocating their food budget between the healthy and the unhealthy categories. A shift out of the budget constraint, keeping the relative prices of different goods fixed, leads to an increase in the consumption level of all normal goods, and a decrease in consumption of inferior goods. Hence, this comparison suggests that the healthy category is relatively normal – and the unhealthy category, relatively inferior.[51] Analogous figures by food product – showing graphically the estimates from Table 10, – can be found in Appendix A.11.3.

I measure the overall effect on nutrition of this change in disposable income as above: I normalize the response to this shift in prices by converting the counterfactual shares into the implied quantities by fixing expenditure on food products, $\sum_{j=1}^{11} P_j q_j$, at a hundred dollars for all households and construct the predicted HEI. The average counterfactual HEI at prices and preferences of the lowest income group is plotted in Figure 9. Each point (x, y) on the light gray curve represents the counterfactual HEI that households from the lowest income group would have, y, if their food budget was increased to the level of income group x. This implies that the lowest income households, upon a increase in the food budget to the level of the highest income households, would change the composition of their food baskets, so that their HEI would almost match the HEI of the highest income households.

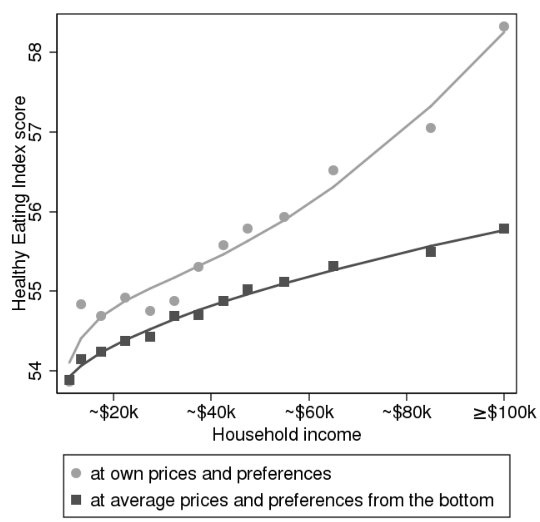


Figure 9: *Average predicted Healthy Eating Index by income group at prices and preferences from the lowest income group in dark gray. Each pair (x, y) on the dark gray curve represents the counterfactual HEI that households from the lowest income group would have, y, if their food budget was increased to the level of income group x.*

I find that the change in overall quantity purchased of healthy and unhealthy products translates into a two-point increase in the nutrient score of low-income households, closing

[51] I use the terms *relatively normal* and *relatively inferior* to emphasize that I refer to normal and inferior goods with respect to food budget rather than to income. As can be verified in Table 2, all food categories are necessities.

over 40% of the gap with high-income households. The rise in the nutrient score reflects changes in the components of the index. Two of the components are grams of protein and calories out of sugar, plotted in Figure 10. The effect of increasing the budget of the lowest income households is represented by the light gray curve. Increasing the budget of the lowest income group to the level of the highest income group, leads to an increase of protein consumption from 23.5 grams to the dashed lined at 24.5 grams (a 4.3% increase). Instead, it lowers the calories out sugar from 139 to the dashed line at 100 kcal. (almost 40% decrease).

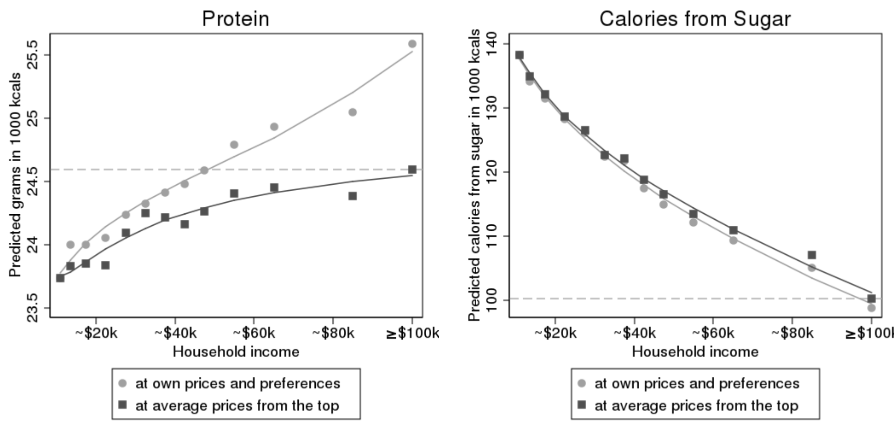


Figure 10: *Average predicted protein and calories from sugar by income group at prices and preferences from the lowest income group in light gray. The dark gray curve plots the predicted protein and calories from sugar at own preferences and prices. The vertical distance between the point corresponding to the lowest income group (where the two curves cross) and the dashed line represents the counterfactual quantity (in grams on the left panel and in calories on the right panel) that households from the lowest income group would purchase if their food budget was increased to the level of the highest income group.*

6.2.3 Preference Effects

The last counterfactual exercise I consider consists on equalizing the preference parameters to the level of the highest income group. That is, I consider

$$\mathrm{w}_{ijt}(I) - \mathbb{E}(\mathrm{w}_{ijt} \mid I_{it} = I, \xi_j^H) = \xi_{ij}(I) - \xi_j^H. \tag{6}$$

The right-hand side of (6) captures the effect on i's consumption of category j of "changing" his preference for j. Again, bv

I summarize the change in composition of the food baskets of consumers by comparing the ration of quantities bought of healthy products with respect to quantities bought of unhealthy products as plotted in Figure 11, which shows that this ratio would increase the ratio of quantities consumed of the healthy category for all income groups. The average of these differences for each income group and each product can be found in Appendix A.10.4.

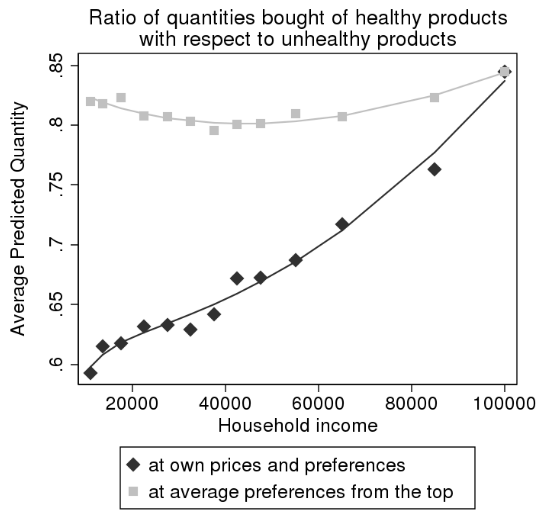


Figure 11: *Predicted ratio of overall quantities purchased (measured in grams) of healthy to unhealthy products by income group at own preferences and at average preferences from the highest income group. The comparison illustrates the effect of equalizing preferences at the average level of the highest income group on the composition food baskets. Appendix A.10.4 shows the effect product by product.*

I compare the Healthy Eating indices obtained using the implied quantities fixing expenditures for all households at the same level when preferences are held at the level of the top income group, $\mathbb{E}(\mathrm{w}_{ijt} \mid I_{it} = I, \xi_i^H)$. The average of these indices by income group is plotted in figure Figure 12.

The difference in scores between the highest and lowest income groups decreases by approximately 44% when the preferences are held at the average level from the top income group.

The overall contribution of prices, preferences and budget to the HEI can be broken down through the following decomposition. Let $H(y_{it}, P_{ijt}, \xi_{ijt})$ define the HEI of household i with his corresponding expenditure, price schedule and preferences. The darker curves in figures 7, 12 and 9 plot the averages $\frac{1}{|I|}\sum_{i \in I} H(y_{it}, P_{ijt}, \xi_{ijt})$ for each income group I, which I denote by $H(y^I, P^I, \xi^I)$.

$$\log\left(\frac{H(y^H, P^H, \xi^H)}{H(y^L, P^L, \xi^L)}\right) = \log\left(\frac{H(y^H, P^L, \xi^L)}{H(y^L, P^L, \xi^L)}\right) + \log\left(\frac{H(y^H, P^H, \xi^L)}{H(y^H, P^L, \xi^L)}\right)$$
$$+ \log\left(\frac{H(y^H, P^H, \xi^H)}{H(y^H, P^H, \xi^L)}\right)$$

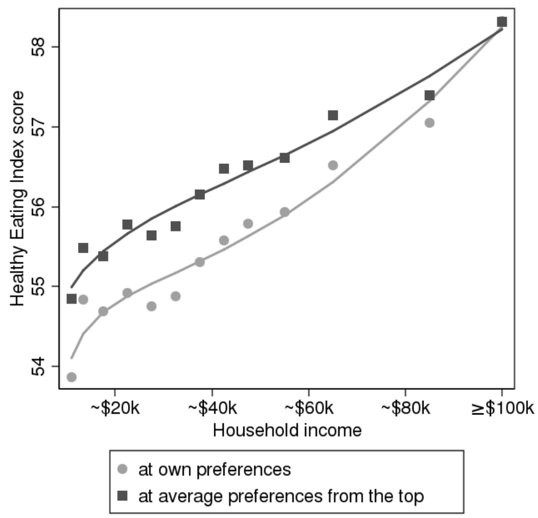


Figure 12: *Average predicted Healthy Eating Index by income group at own preferences and at average preferences from the highest income group. The comparison illustrates the effect on the nutrient scores of equalizing preferences at the average level of the highest income group.*

where the the first term in the right hand side measures the contribution of the preference effect, the second term captures the price effect, and the third term captures the effect of a shift in budget – as that the only difference between the bottom and top group once preferences and prices are equalized. These terms equal, respectively, 0.0150, 0.0032 and 0.0160. That is the preference contribution to the gap between $H(y^H, P^H, \xi^H)$ and $H(y^L, P^L, \xi^L)$ is 43.88%, the price contribution is 9.31%, and the budget has a contribution of 46.82%.

6.3 Preference Parameters and Demographics

Preferences for healthy products could be either innately different for different income groups, or shaped by long-term exposure to characteristics that are linked to own-income or to the income level of the households' environment. In this section I show that irrespective of the income level of their zipcode, their race and the hours of employment of the heads of the household, households with higher income have higher preferences for healthful products, suggesting that own-income plays a larger role than the environment of the household. Moreover, conditional on income, households with more educated heads of the household tend to have higher preferences for healthful foods.

In Appendix A.12 I show the correlation of preferences with other variables product by product. Overall, I find that income is positively correlated with preferences for healthy products (this is consistent with the findings of Allcott et al. (2017) and Handbury (2013)).

Moreover, conditional on income level, more educated households tend to have higher preferences for healthier goods. Households with higher socioeconomic characteristics tend to have higher taste for healthier food groups, in particular as measured by education. To see this, I regress the estimated preference parameters on education, and income quartile dummies and their interactions, together with zipcode-level-income fixed effects. The estimated parameters from these regressions are reported in Table 16a in Appendix A.12. Adding dummies for race and number of employment hours yields similar estimates.

I define the preference for the healthy category g_i^h as the sum of the preference parameters for the products that are classified as healthy with respect to the sum of all preferences for food categories, $g_i^h = \frac{\sum_{c \in H} \xi_{ic}}{\sum_{c'} \xi_{ic'}}$, to summarize the association of preferences for healthy products with different demographic and socioeconomic characteristics. The correlation of these preferences for healthy products separately by education group at each income level are plotted in Figure 13 (and similar figures by product can be found in Appendix A.12.)

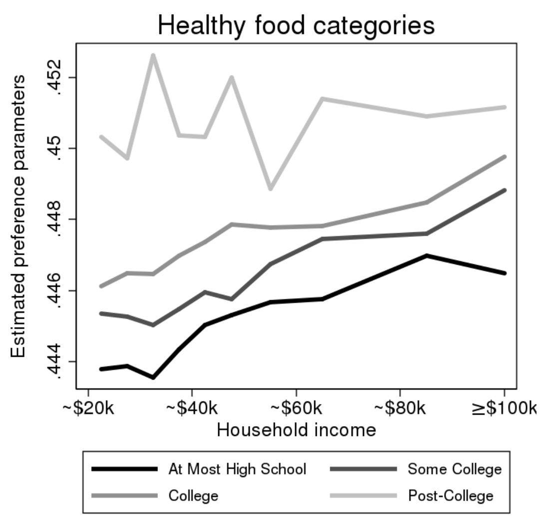


Figure 13: *Correlation of these preferences separately by education group at each income level.*

Healthier products, such as vegetables and fresh meat require more preparation time than frozen, canned and prepared food. Thus, longer employment hours may induce unhealthier consumption.[52]

Figure 14 plots the correlation of preferences for healthful products separately by employment hours of the male head by household income. Although the correlation between preferences for healthy products and range of hours worked is consistent with this hypothesis, income seems to play a larger role.

[52]Note that non-market (or leisure) time available for activities such as shopping for goods and food preparation is not increasing in income. In Appendix A.5 I tabulate time-use by income quintile.

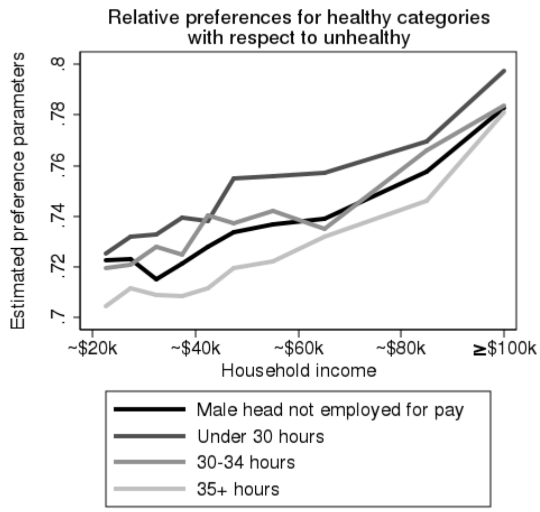


Figure 14: *Correlation of preferences for healthy products separately by number of employment hours of the heads at each income level.*

I find that the preference parameters correlate with reported race which can be correlated with the cultural identity of individuals. If race was in fact correlated with cultural identity, the fact that households of different phenotype display different consumption patterns would align with preferences being sticky (Atkin (2016)). Unfortunatedly, the Kilts Center does not collect any variable related to country of origin in the Nielsen Homescan data. Hence I cannot further explore this dimension.[53] Asians in particular seem to have higher taste for healthier products (Figure 15).

Finally, I use the ACS to group zipcodes by average household-level income. Ranking zipcodes by their average income level, I can look at the correlation between preferences for healthy products by own-household income separately for different zipcode-levels of income. I find that conditional on income, zip-level income correlates with preferences.

[53]There are also significant differences in preferences by Hispanic origins. However, only 5% of the sample resports having Hispanic origins.

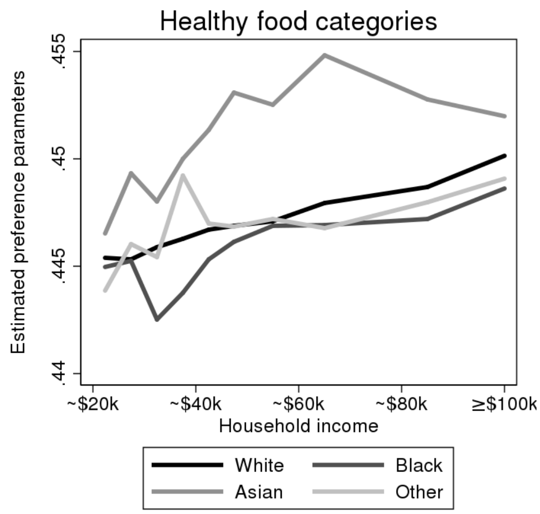


Figure 15: *Correlation of preferences for healthy products separately for each racial group at each income level.*

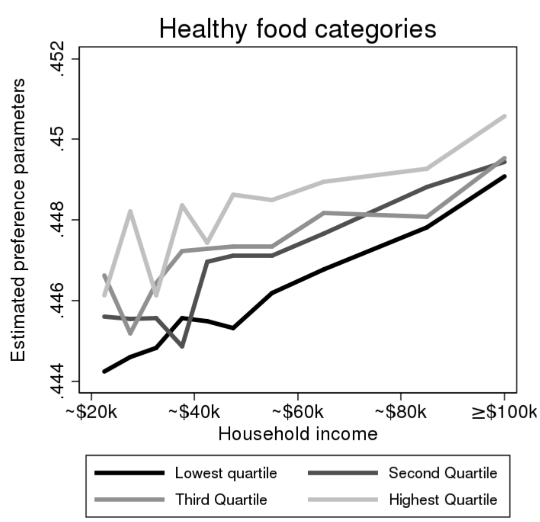


Figure 16: *Correlation of preferences for healthy products separately zip-level income at each income level.*

7 Conclusion

This paper studies the contribution of prices, food expenditure and preferences to the differences in consumption patterns across income groups. I estimate a demand system for different food products using longitudinal variation, which allows me to disentangle the direct income and price effects from preferences.

While previous research has compared the consumption baskets of households at different income levels, I exploit within and across household variation in the level and composition of food expenditure. This source of variation allows me to disentangle the direct income effect – e.g., a positive shock in current disposable income, – from the effect of permanent differences across individuals (or preferences.)

Performing a series of counterfactual exercises with the structural estimates of the model and making use of nutritional data to map changes in consumption into a nutrition score, I find that disposable income and preferences have a predominant and quantitatively similar role in explaining food consumption differences across income groups. Instead, prices have limited effects since low-income households face relatively lower prices for healthy goods.

The pure income effect and the time invariant preference parameters speak to different aspects of policy development. The pure income effect is relevant for programs that are designed to shift the food budget of households. The preference parameters, instead, inform policies targeted at shaping the environment of different socioeconomic groups. My results suggest that policies that increase the food budget of low-income families can have a positive effect on nutrition quality. A limitation of the results is that they only speak to policies that affect the disposable income for food products.[54] Moreover, even with extreme counterfactual policies in which the entire gap in food budget is closed, differences in consumption remain due to differences in preferences.

I take a first step in uncovering the nature of preferences by analyzing their association with socioeconomic and demographic characteristics. I find that irrespective of the income level of their zipcode, their race and the hours of employment of the heads of the household, households with higher income have higher preferences for healthful products, suggesting that own-income plays a larger role than the environment of the household. Moreover, conditional on income, households with more educated heads of the household tend to have higher preferences for healthful foods. These associations between preferences and socioeconomic characteristics hint to a nuanced relationship between income and preferences and suggest that a natural next step in explaining the drivers of nutrition inequality is to formally analyze

[54]Recent literature analyzing the marginal propensity to consume food out of food stamp benefits, suggests that this program succeeds at increasing food budget.

the components of the time-invariant preference parameters. In particular, an interesting question for future research is whether one could design a policy intervention that mitigates the differences in preferences for food.

References

Allcott, Hunt, Rebecca Diamond, and Jean-Pierre Dube (2017) "The Geography of Poverty and Nutrition: Food Deserts and Food Choices Across the United States," NBER Working Paper No. 24094.

Allcott, Hunt, Rebecca Diamond, Jean-Pierre Dube, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell (2018) "Food Deserts and the Causes of Nutritional Inequality," NBER Working Paper No. 24094.

Altonji, Joseph G., Anthony A. Smith, and Ivan Vidangos (2013) "Modeling Earnings Dynamics," *Econometrica*, 81, 1395–1454.

Atkin, David (2016) "The Caloric Costs of Culture: Evidence from Indian Migrants," *The American Economic Review*, 106, 1144–1181.

Banks, James, Richard Blundell, and Arthur Lewbel (1997) "Quadratic Engel Curves and Consumer Demand," *Review of Economics and Statistics*, 79, 527–539.

Broda, Christian, Ephraim Leibtag, and David E. Weinstein (2009) "The Role of Prices in Measuring the Poor's Living Standards," *The Journal of Economic Perspectives*, 23, 77–97.

Bruich, Gregory A. (2014) "The Effect of SNAP Benefits on Expenditures: New Evidence from Scanner Data and the November 2013 Benefit Cuts," Harvard University, Working Paper.

Chernozhukov, Victor and Han Hong (2003) "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293–346.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler (2016) "The Association between Income and Life Expectancy in the United States, 2001-2014," *The Journal of the American Medical Association*, 315, 1750–1766.

Claro, Rafael Moreira, Renata Bertazzi Levy, Daniel Henrique Bandoni, and Lenise Mondini (2010) "Per Capita versus Adult-Equivalent Estimates of Calorie Availability in Household Budget Surveys," *Cadernos de Saude Publica*, 26, 2188–2195.

Cummins, Steven, Ellen Flint, and Stephen A. Matthews (2014) "New Neighborhood Grocery Store Increased Awareness of Food Access but did not Alter Dietary Habits or Obesity," *Health Affairs*, 33, 283–291.

Deaton, Angus (1988) "Quality, Quantity, and Spatial Variation of Price," *The American Economic Review*, 418–430.

Deaton, Angus and John Muellbauer (1980) "An Almost Ideal Demand System," *The American Economic Review*, 70, 312–326.

Dubois, Pierre, Rachel Griffith, and Aviv Nevo (2014) "Do Prices and Attributes Explain International Differences in Food Purchases?" *The American Economic Review*, 104, 832–867.

Dubowitz, Tamara, Madhumita Ghosh-Dastidar, Deborah A. Cohen, Robin Beckman, Elizabeth D. Steiner, Gerald P. Hunter, Karen R. Flórez, Christina Huang, Christine A. Vaughan, and Jennifer C. Sloan (2015) "Diet and Perceptions Change with Supermarket Introduction in Food Desert, but not Because of Supermarket Use," *Health Affairs*, 34, 1858–1868.

Dustmann, Christian and María Engracia Rochina-Barrachina (2007) "Selection Correction in Panel Data Models: An Application to the Estimation of Females' Wage Equations," *The Econometrics Journal*, 10, 263–293.

Einav, Liran, Ephraim Leibtag, and Aviv Nevo (2008) "Not-so-Classical Measurement Errors: a Validation Study of Homescan," National Bureau of Economic Research, Working Paper No. 14436.

Fally, Thibault and Benjamin Faber (2016) "Firm Heterogeneity in Consumption Baskets: Evidence from Home and Store Scanner Data," National Bureau of Economic Research, Working Paper No. 23101.

Gorman, Willam M. (1996) "Two Stage Budgeting," in C. Blackorby and A. F. Shorrocks eds. *Separability and Aggregation: The Collected Works of W. M. Gorman, Volume I*, Oxford, Oxford University Press, Chap. 2.

Gourieroux, C., A. Monfort, and E. Renault (1993) "Indirect inference," *Journal of Applied Econometrics*, 8, S85–S118.

Handbury, Jessie (2013) “Are Poor Cities Cheap for Everyone? Non-Homotheticity and the Cost of Living across U.S. Cities,” National Bureau of Economic Research, Working Paper No. 18314.

Handbury, Jessie, Ilya Rahkovsky, and Molly Schnell (2015) “Is the focus on food deserts fruitless? Retail access and food purchases across the socioeconomic spectrum,” National Bureau of Economic Research, Working Paper No. 21126.

Hastings, Justine S. and Jesse M. Shapiro (2017) “How Are SNAP Benefits Spent? Evidence from a Retail Panel,” National Bureau of Economic Research.

Hausman, Jerry A. (1996) “Valuation of New Goods under Perfect and Imperfect Competition,” in *The Economics of New Goods*, University of Chicago Press, 207–248.

Jaravel, Xavier (2016) “The Unequal Gains from Product Innovations: Evidence from the US Retail Sector,” Stanford Institute for Economic Policy Research, Working Paper.

Kaplan, Greg and Guido Menzio (2016) “Shopping Externalities and Self-Fulfilling Unemployment Fluctuations,” *Journal of Political Economy*, 124, 771–825.

Kozlova, Olga (2016) “Is Healthy Food a Luxury for the Low-Income Households in the US?” Duke University, Working Paper.

Lee, Lung-Fei and Mark M. Pitt (1986) “Microeconometric Demand System with Binding Nonnegativity Constraints: the Dual Approach,” *Econometrica*, 1237–1242.

Lewbel, Arthur and Krishna Pendakur (2009) “Tricks with Hicks: The EASI Demand System,” *The American Economic Review*, 99, 827–863.

Micha, Renata, Jose L. Peñalvo, Frederick Cudhea, Fumiaki Imamura, Colin D. Rehm, and Dariush Mozaffarian (2017) “Association Between Dietary Factors and Mortality From Heart Disease, Stroke, and Type 2 Diabetes in the United States,” *The Journal of the American Medical Association*, 317, 912–924.

Mozaffarian, Dariush, Renata Micha, and Sarah Wallace (2010) “Effects on Coronary Heart Disease of Increasing Polyunsaturated Fat in Place of Saturated Fat: A Systematic Review and Meta-Analysis of Randomized Controlled Trials,” *PLoS Medicine*, 7.

Neary, J. Peter and Kevin W.S. Roberts (1980) “The Theory of Household Behaviour under Rationing,” *European Economic Review*, 13, 25–42.

Olmos, Pablo (2017) "Time Poverty in Developing Countries," Yale University, Available at https://pabloolmos.wordpress.com/research/.

Phaneuf, Daniel J. (1999) "A Dual Approach to Modeling Corner Solutions in Recreation Demand," *Journal of Environmental Economics and Management*, 37, 85–105.

Wales, Terence J. and Alan Donald Woodland (1983) "Estimation of Consumer Demand Systems with Binding Non-negativity Constraints," *Journal of Econometrics*, 21, 263–285.

Zhen, Chen, Eric A. Finkelstein, James M. Nonnemaker, Shawn A. Karns, and Jessica E. Todd (2013) "Predicting the Effects of Sugar-sweetened Beverage Taxes on Food and Beverage Demand in a Large Demand System," *American Journal of Agricultural Economics*, 96, 1–25.

A Appendix

A.1 Food Consumption and Nutrient Intake

Table 3: *Average daily intake of food by food source and demographic characteristics, 2007-10*

Food group	**Total**	**At home**	**Away from home**				
			Total	Restaurant	Fast food	School	Other
Fruits (cups)							
Total population	1.05	0.94	0.12	0.02	0.02	0.03	0.05
Children	1.08	0.90	0.18	0.01	0.02	0.10	0.05
Adults	1.05	0.95	0.10	0.02	0.02	0.00	0.06
Lower income	1.01	0.89	0.12	0.02	0.01	0.04	0.05
Higher income	1.08	0.96	0.11	0.02	0.02	0.02	0.05
Vegetables: total (cups)							
Total population	1.42	0.96	0.47	0.18	0.17	0.02	0.10
Children	0.92	0.59	0.33	0.07	0.14	0.07	0.05
Adults	1.59	1.08	0.51	0.21	0.18	0.01	0.11
Lower income	1.26	0.88	0.38	0.11	0.16	0.03	0.08
Higher income	1.53	1.01	0.52	0.22	0.17	0.02	0.11
Dark green vegetables (cups)							
Total population	0.11	0.08	0.03	0.02	0.01	0.00	0.01
Children	0.05	0.04	0.01	0.01	0.00	0.00	0.00
Adults	0.14	0.10	0.04	0.02	0.01	0.00	0.01
Lower income	0.08	0.07	0.02	0.01	0.00	0.00	0.01
Higher income	0.13	0.10	0.04	0.02	0.01	0.00	0.01
Red and orange vegetables (cups)							
Total population	0.37	0.26	0.11	0.04	0.04	0.01	0.02
Children	0.27	0.17	0.09	0.02	0.04	0.02	0.02
Adults	0.40	0.28	0.12	0.05	0.05	0.00	0.03
Lower income	0.33	0.24	0.09	0.03	0.04	0.01	0.02
Higher income	0.39	0.27	0.13	0.05	0.05	0.01	0.03
Tomatoes (cups)							
Total population	0.28	0.19	0.10	0.03	0.04	0.01	0.02
Children	0.22	0.13	0.08	0.02	0.04	0.02	0.01
Adults	0.31	0.21	0.10	0.04	0.04	0.00	0.02
Lower income	0.26	0.18	0.08	0.02	0.04	0.01	0.01
Higher income	0.30	0.19	0.11	0.04	0.04	0.00	0.02
Dairy (cups)							
Total population	1.77	1.32	0.45	0.10	0.19	0.08	0.08
Children	2.16	1.52	0.64	0.08	0.20	0.29	0.08
Adults	1.64	1.26	0.38	0.11	0.19	0.01	0.08
Lower income	1.67	1.26	0.41	0.06	0.18	0.11	0.06
Higher income	1.83	1.36	0.47	0.13	0.20	0.06	0.09

Food group	Total	At home	Away from home				
			Total	Restaurant	Fast food	School	Other
Refined grains (ounces)							
Total population	5.68	3.75	1.93	0.51	0.89	0.13	0.40
Children	5.84	3.87	1.97	0.30	0.84	0.48	0.35
Adults	5.63	3.71	1.92	0.58	0.91	0.02	0.41
Lower income	5.67	3.90	1.77	0.37	0.87	0.18	0.35
Higher income	5.69	3.65	2.04	0.60	0.90	0.10	0.43
Whole grains (ounces)							
Total population	0.78	0.72	0.06	0.02	0.01	0.01	0.03
Children	0.62	0.56	0.06	0.01	0.00	0.03	0.02
Adults	0.83	0.77	0.07	0.02	0.01	0.00	0.03
Lower income	0.67	0.62	0.05	0.01	0.01	0.01	0.02
Higher income	0.85	0.77	0.07	0.02	0.01	0.01	0.03
Protein foods (ounces)							
Total population	5.68	3.83	1.85	0.65	0.72	0.09	0.39
Children	4.33	2.88	1.45	0.28	0.63	0.31	0.22
Adults	6.13	4.15	1.98	0.77	0.75	0.02	0.44
Lower income	5.29	3.67	1.62	0.45	0.71	0.12	0.33
Higher income	5.93	3.94	1.99	0.77	0.73	0.07	0.42
Added sugars (tsp)							
Total population	17.73	13.17	4.56	0.98	1.59	0.28	1.72
Children	18.43	13.17	5.26	0.87	1.70	0.99	1.70
Adults	17.50	13.17	4.33	1.01	1.55	0.04	1.73
Lower income	18.59	14.26	4.32	0.78	1.67	0.38	1.51
Higher income	17.18	12.48	4.71	1.10	1.54	0.21	1.86
Oils (grams)							
Total population	21.15	13.75	7.40	2.38	3.16	0.39	1.47
Children	17.97	11.30	6.67	1.17	3.19	1.33	0.98
Adults	22.20	14.56	7.64	2.78	3.15	0.08	1.64
Lower income	19.12	12.84	6.28	1.56	3.10	0.48	1.15
Higher income	22.43	14.32	8.11	2.90	3.20	0.33	1.68
Solid fats (grams)							
Total population	37.44	24.44	13.00	3.53	5.51	0.88	3.08
Children	36.08	23.29	12.79	2.02	5.13	3.18	2.47
Adults	37.89	24.82	13.07	4.03	5.64	0.12	3.28
Lower income	35.72	24.09	11.63	2.48	5.48	1.21	2.46
Higher income	38.53	24.67	13.86	4.19	5.53	0.67	3.47

* **Source:** This information was obtained through the USDA *Food Consumption and Nutrient Intake* data (https://www.ers.usda.gov/Data/FoodConsumption/).
These average intakes are based on the National Health and Nutrition Examination Survey (NHANES) for 2007-2010 using two-day averages.
Individuals who completed two-day intake recalls, were 2 years and older, and were not pregnant or lactating. Children are those age 2-19 and adults are age 20 and older. The 185 percent income poverty line is the income threshold separating lower and higher income households.

A.2 Details about the Data

A.2.1 Nielsen Homescan Data

The Nielsen Homescan Data provides a detailed product hierarchy. About 3 million products (identified by their barcode, or UPC code) are classified into 10 broad departments, from which I focus on those related with food (in particular, Dry grocery, Frozen Foods, Dairy, Deli, Packaged Meat, Fresh Produce); this reduces the number of barcodes to 2 million. Nielsen defines two more levels of hierarchy.

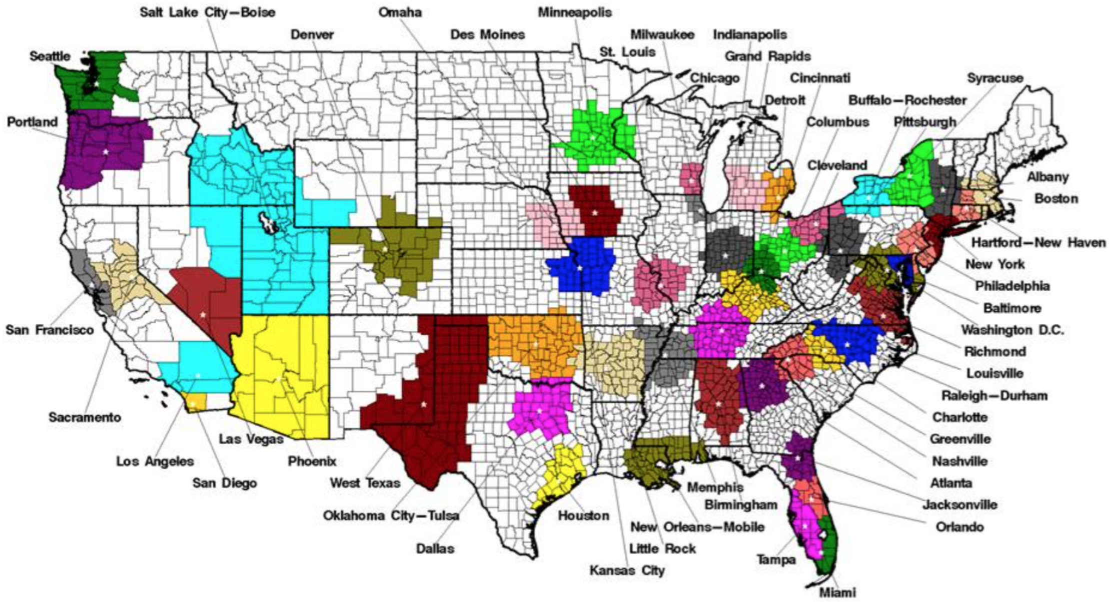
Nielsen data does not contain actual annual household income. Instead, panelists select one of the following ranges in which their income falls: less than \$5000, \$5000-\$7999, \$8000-\$9999, \$10,000-\$11,999, \$12,000-\$14,999, \$15,000-\$19,999, \$20,000-\$24,999, \$25,000-\$29,999, \$30,000-\$34, 999, \$35,000-\$39,999, \$40,000-\$44,999, \$45,000-\$49,999, \$50,000-\$59,999, \$60,000-\$69,999, \$70,000-\$99,999 and over \$100,000 (for most years).[55] The bottom four income groups (up to the \$11,999 limit) represent less than 5% of the sample each year and their probability weights are, on average, almost double the ones for individuals with income ranges between \$12,000 and \$100,000. Thus I group these bottom four income groups together to analyze consumption patterns by income.[56]

Nielsen selects panelists trying to match the demographics of the U.S. population on household size, household income, age of the head of household, race, Hispanic origin, education and occupation of the head of the household, presence of children and Nielsen county size. The sample can be analyzed at the level of *Scantrack markets*, which typically consist of a set of adjacent counties and are shown in Figure 17.

[55] For 2006-2009, the upper ranges are instead \$100,000 - \$124,999, \$125,000 - \$149,999, \$150,000 - \$199,999, and over \$200,000.

[56] Household income is recorded two years prior to the panel year. Panelists are asked in the fall prior to the start of the panel year for their total annual income as of year-end of the previous calendar year. I use the income information reported each year as current income following the literature that uses income information from the Nielsen Homescan Data (see Jaravel (2016), Fally and Faber (2016).)

Figure 17: Nielsen total U.S. Scantrack markets.



The subsample of households who report random weighted items, however, can only be analyzed at the level of the country.

A.2.2 Gladson Nutritional Data

The Gladson Nutrition Database includes detailed nutritional attributes of thousands of food products. Unfortunately, it only covers a fraction of the products registered by the respondents of the Nielsen Homescan Data.

The linkage is further complicated by the fact that the nutrients in the Gladson Data are measured as a function of serving size. This serving size is, in many cases, measured in unconventional units (such as "one 4-inch diameter cooked pancake"), which combined with the large fraction of products that have a vague number of servings per container (some have "varies" as the number of servings per container) makes many of the matched barcodes' nutritional contents futile.

Once I restrict the observations in the Gladson data to those for which I can redefine the serving size in metric units, the matching rate by year is as shown below.

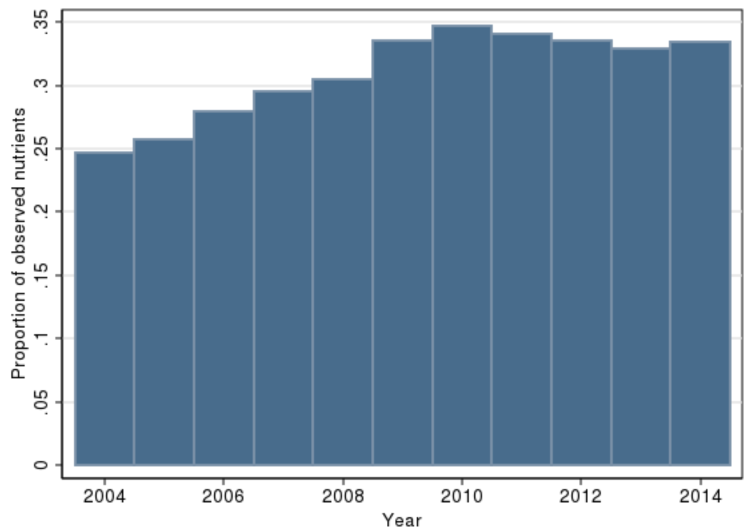


Figure 18: *Proportion of products in the Nielsen Homescan Data for which I can merge their corresponding nutritional contents from the Gladson Data.*

Fortunately, the matching rate is similar across income groups.

Out of the total of 1,692,946 barcodes in the food categories considered, 14.12% is matched directly.

Using the products' brand and attributes – including the description of the product (for example "canned tomato, whole") and characteristics such as organic claim, salt content and the "common consumer name description" and "variety description" (when available) in the Nielsen Homescan Data, I grouped barcodes without nutritional information into detailed product groups. I used analogous information in the Gladson nutritional data (based on the brand, organic claim, sodium content, and product description) and matched the resulting product groups. This method allows me to project this information to 61.11%. Using all the characteristics but brand, I regroup the barcodes with missing information in both datasets and match 42.71% of the remaining products without nutritional information (this yields a total of 87.96% of products with nutrients).

A.3 The Role of Prices in Nutritional Disparities

Lower income households pay relatively lower prices for healthy products than for unhealthy ones. To see this, I define price indices to compare the prices for products in the healthy category with respect to those in the unhealthy category. I use a weighted average of the actual prices paid for a product, with weights reflecting the product shares at the national level for a given year. In particular, let j denote food categories, i the households in the sample and t a year. I define P_{ijt} to be

$$P_{ijt} = \sum_{u \in U_{jt}(i)} \frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}} P_{it}(u)$$

with $U_{jt}(i)$ the set of barcodes u in category j that i bought in t and $\frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}}$ defined as the (nation-wide) expenditure share on barcode u – these are the price indices I use for estimation (where shorten t to half year periods).

In Figure 19, I plot the ratio of the average price indices in the healthy category with respect to the unhealthy category.

$$\rho_i = \frac{1}{|Healthy|} \sum_{j \in Healthy} P_{ijt} \Big/ \frac{1}{|Unhealthy|} \sum_{j \in Unealthy} P_{ijt}.$$

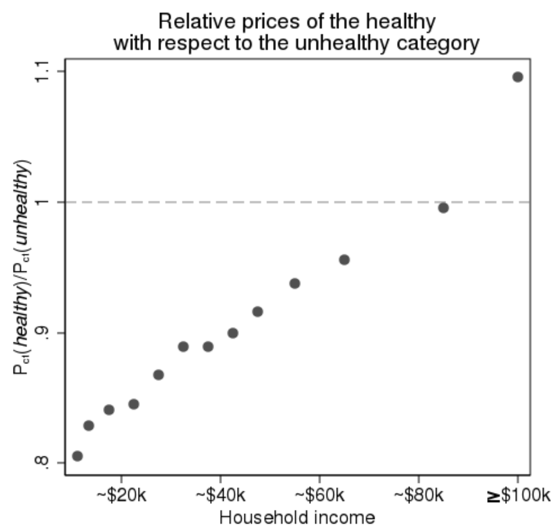


Figure 19: *Average relative prices paid for products in the healthy category with respect to products in the unhealthy category by income group.*

Note that ρ_i is below 1 if the price index of healthy products is below that of unhealthy products. Hence, Figure 19 implies that lower income households pay relatively lower prices for the healthy products they buy with respect to the unhealthy products they buy.

A.3.1 Relative Prices of the Healthy and Unhealthy products by income

To compare differences in prices paid across counties, similarly, I define the aggregate price index for county c in year t by

$$\mathbf{P_{cjt}} = \sum_{u \in U_{jt}(c)} \frac{v_{ut}}{\sum_{u \in U_{jt}} v_{ut}} P_{ct}(u)$$

with $U_{jt}(c)$ the set of barcodes u in category j that were bought in county c at t and $\frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}}$ defined as the (nation-wide) expenditure share on barcode u.

I consider, as above, the ratio of healthful-to-unhealthful price indexes

$$\frac{1}{|Healthy|} \sum_{j \in Healthy} P_{cjt} \Bigg/ \frac{1}{|Unhealthy|} \sum_{j \in Unealthy} P_{cjt}.$$

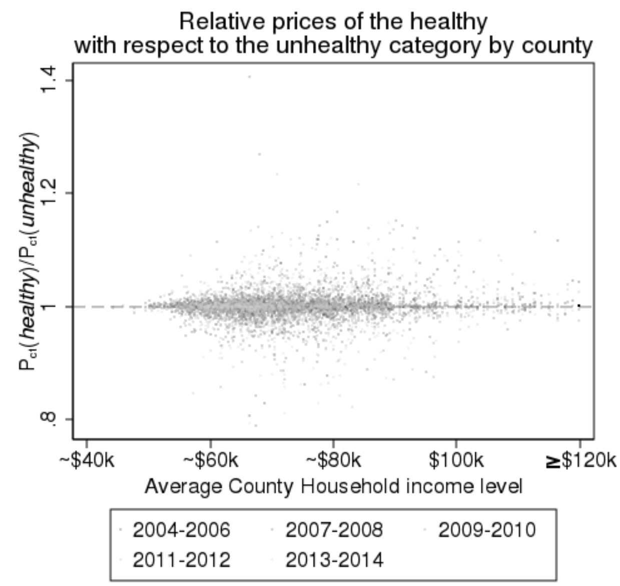
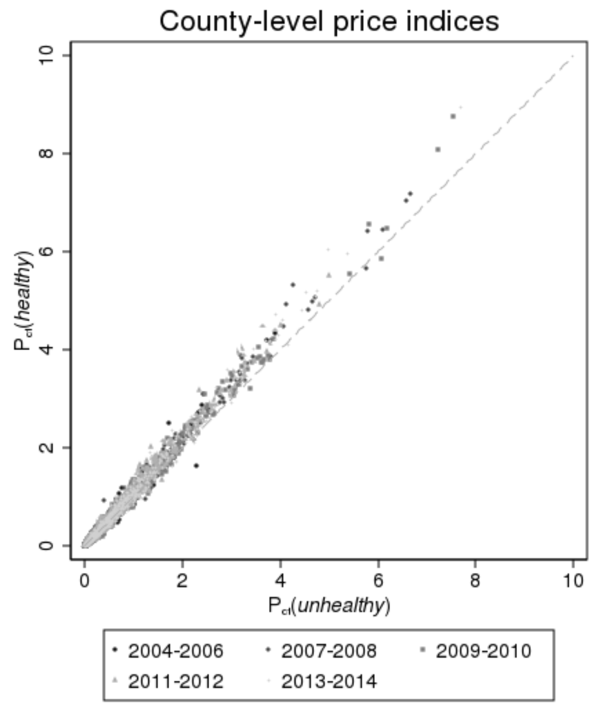


Figure 20: *Counties are ranked by average household income level.*

This suggests that the relative price indexes of the healthy to the unhealthy products are not strongly correlated with county-level demographics.

A plot of P_{cjt}^{healthy} on $P_{cjt}^{\text{unhealthy}}$ yields



where the dashed gray line is the 45-degree line.

This shows that, at the barcode-level, the correlation between county-level household income and prices is weak.

A.3.2 Higher income households buy more expensive products than lower income households

Consider the relative price defined as the coefficient on a household income dummy in a regression of the log unit price paid by a household for a product on food category,[57] fixed effects and demographic controls. That is, let the relative price of a product relative to the category average to be given as the coefficient $\hat{\beta}_y$ in the regression below,

$$\log(p_{iu}) = \sum_{y \in Y} \beta_y \mathbb{1}(income_i = y) + \alpha D_i + \sum_{j \in J} \gamma_j \mathbb{1}(category_u = j) + \varepsilon_{iu}$$

where i indicates household, t a time period, u indicates barcode and D_i denotes a vector of dummies indicating demographic characteristics and $\mathbb{1}(category_u = j)$ are product category controls.

The β coefficients, plotted with light gray squares in Figure 21, tell us that the prices

[57]Where recall, a food category is a group of food products that was found to be either positively or negatively correlated with chronic conditions.

Figure 21: *Relative price defined as the coefficient on a household income dummy in a regression of the log unit price paid by a household for a product on food category fixed effects and demographic controls.*

paid by households for products in a food category are increasing with income.[58] These prices could be higher because higher income households pay more for the same products in a category, or because they buy different products.

But when focusing on a narrower classification of products, namely modules (these are groups of barcodes such as "canned tomato, whole", we see that this is not completely driven by the prices paid by high income households. Instead, the relative prices $\hat{b}_y$ from the regression below,

$$\log(p_{iut}) = \sum_{y \in Y} b_y \mathbb{1}(income_{it} = y) + \alpha D_i + \sum_{j \in J} \delta_j \mathbb{1}(module_u = j) + \varepsilon_{iut}.$$

plotted in medium gray diamonds, show a flatter slope, and the parameters corresponding to the most detailed classification products, that of barcodes, is almost completely flat. This is evidence that part of the positive correlation that we see between prices paid and income at the food category level, is driven by the products chosen within each category by households of different income levels. That is, higher income households buy different and more expensive products than lower income households – consistent with Handbury (2013), Kaplan and Menzio (2016) and Broda et al. (2009).

[58] Statistically insignificant coefficients at the 95% confidence level were replaced by zero.

Moreover, separating products into the healthful and unhealthful categories defined above – and defining $\hat{\beta}_y$ and $\hat{b}_y$ by running the regressions above separately for the healthy and the unhealthy products,– we see that this pattern is more pronounced among the healthy category as seen in Figure 22. Implying that high-income households buy products corresponding to the healthful category from different modules.

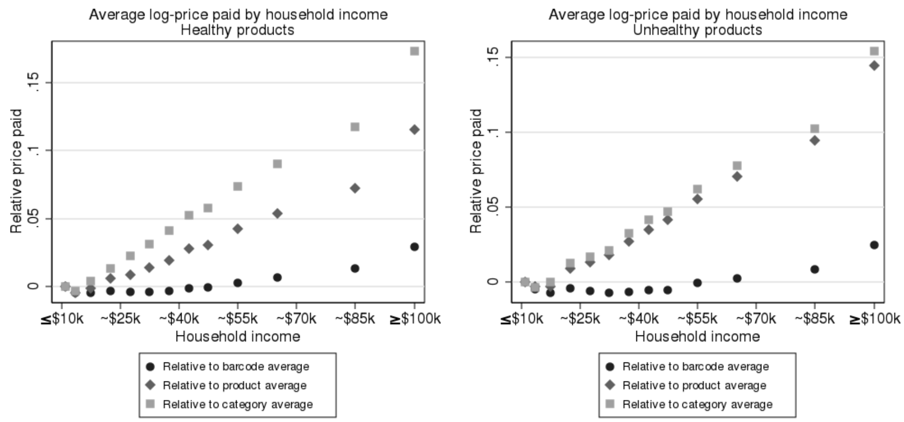


Figure 22: *Relative prices defined as the coefficient on a household income dummy in a regression of the log unit price paid by a household for a product on food category fixed effects and demographic controls, separately by healthful and unhealthful products.*

A.4 Differential Nutrient Consumption

A.4.1 Caloric Composition

Lower income groups consume more saturated fats and – although they source a similar share of their calories out of carbohydrates, – they consume more sugar. The left panel in figure (23) plots the proportion of calories obtained from saturated fats with respect to households from the top income group.[59] The right panel plots the share of saturated fat calories out of the total fat calories.

Similarly, the left panel in figure (24) plots the proportion of calories obtained from sugar with respect to households from the top income group and the right panel the share of calories obtained from sugar out of total calories from carbohydrates.

[59] Saturated fats' consumption is associated with increase in the levels of low-density lipoprotein (LDL) cholesterol which is positively correlated coronary heart disease. There is no consensus in the medical community regarding the effect of saturated fat consumption, however, there is evidence that replacing saturated fats for unsaturated fatty acids lowers LDL cholesterol and increases high-density lipoprotein (HDL) cholesterol that is associated with a reduction in risk for heart disease (Mozaffarian et al. (2010)). Substitution of saturated fats for carbohydrates does not seem to have positive effects. Trans fat consumption has been consistently found to be linked to cardiometabolic diseases (Micha et al. (2017)) however, this type of fat is not reported consistently in the Gladson Nutritional Data.

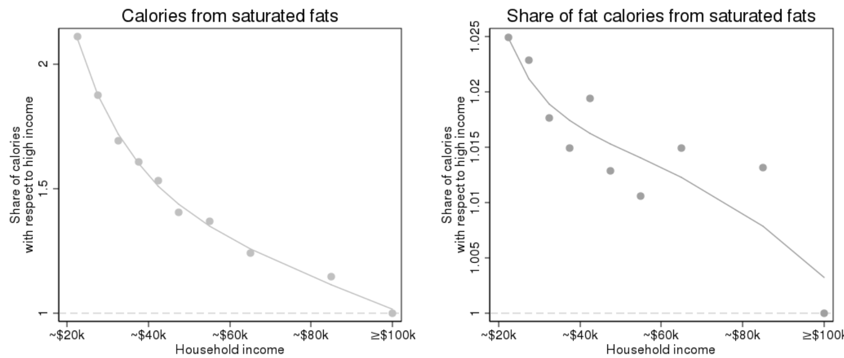


Figure 23: *Proportion of calories obtained from saturated fats with respect to the top income group.*

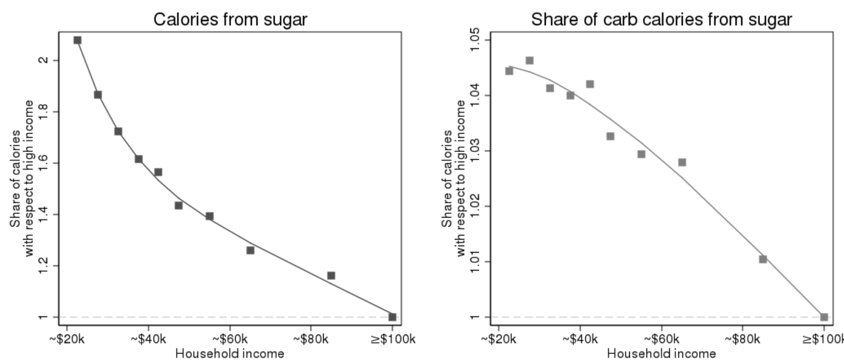


Figure 24: *Proportion of calories obtained from sugar with respect to the top income group.*

A.4.2 Components of the Healthy Eating Index

The Healthy Eating Index is a score based on quantities per 1000 kcal. consumed of different food groups.

To take this measure to the data, I grouped together the components corresponding to fruits and vegetables into the category "fruits and vegetables." Similarly, I use nutritional information on proteins rather than separately quantify the consumption of total protein foods and seafood and plant proteins.

Table 4: *Components of the Healthy Eating Index.*

Component		**Max. points**	**Std for Max. Points**	**Std for zero points**
Fruits and Vegetables		20	$\geq$ 2.5 cups	0
Grains		10	$\geq$ 1.5 ounce	0
Dairy		10	$\geq$ 1.3 cup	0
Protein		10	$\geq$ 3.3 ounce	0
Fatty Acids		10	1.2 $\leq$ (fat - satfat)/satfat $\leq$ 2.5	0
Refined Grains		10	$\leq$ 1.8 oz	$\geq$ 4.3 oz
Sodium		10	$\leq$ 1.1 gram	$\geq$ 2.0 grams
Empty Calories				
	Sugar Calories	10	$\leq$ 6.5% of energy	$\geq$ 26% of energy
	Saturated Fat Calories	10	$\leq$ 8% of energy	$\geq$ 16% of energy

* **Source:** This information was obtained through the USDA *Food Consumption and Nutrient Intake* data (https://www.cnpp.usda.gov/healthyeatingindex).

Note: For consumption levels between the standard for maximum points and the standard for zero points, I do a linear interpolation to assign a number of points between 0 and the maximum attainable points.

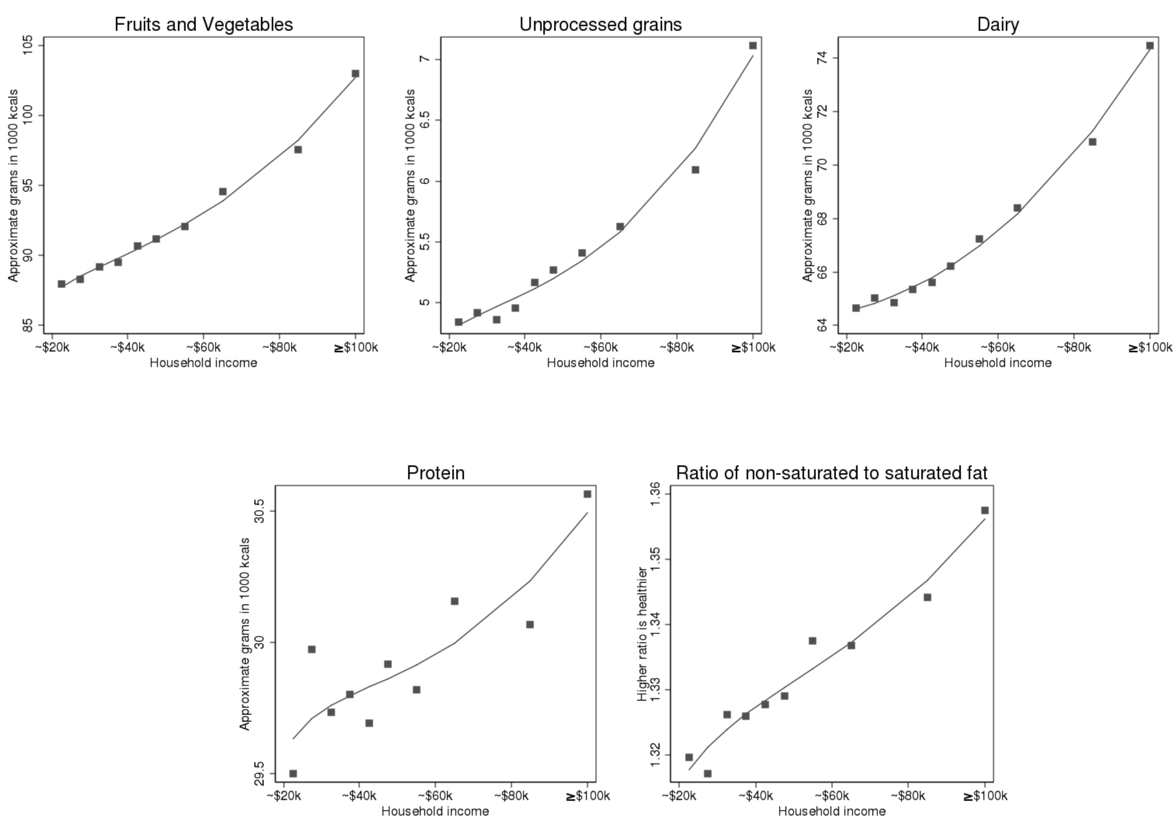


Figure 25: *Average consumption in grams per 1000 kcal. of dietary components that are assigned more points the higher the consumption.*

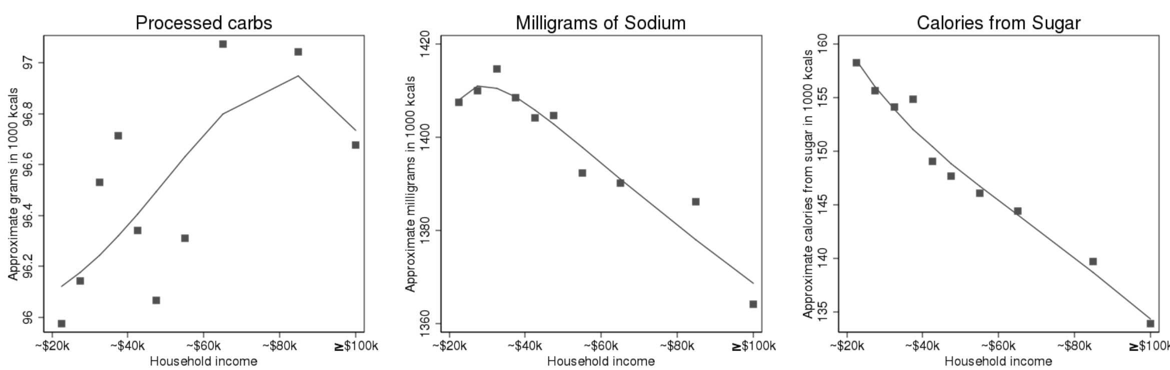


Figure 26: *Average consumption in grams per 1000 kcal. of dietary components that are assigned less points the higher the consumption.*

A.5 Time use by Income Group

Time-use category (hours per week)	USA (2005-2007)				
	Q1	Q2	Q3	Q4	Q5
Panel 1: Market work	27.78	35.46	36.08	38.74	38.01
Core market work	25.05	32.56	33.00	35.27	34.51
Commuting	2.08	2.47	2.69	3.04	3.15
Other income activities	0.22	0.13	0.15	0.14	0.11
Panel 2: Nonmarket work	18.75	18.43	19.42	18.70	18.98
Home production	12.23	11.32	11.81	11.09	10.87
Obtaining goods and services/shopping	4.93	5.19	5.45	5.51	5.97
Garden & Pet	1.59	1.92	2.16	2.10	2.15
Panel 3: Care for others	6.52	6.50	7.23	7.19	7.57
Child care	5.27	5.33	6.05	6.30	6.80
Other care	1.25	1.17	1.17	0.89	0.77
Panel 4: Leisure	37.38	33.59	32.01	30.28	30.13
TV	20.70	16.83	15.27	13.57	11.92
Socializing	7.73	7.00	7.14	6.54	6.69
Exercise & Sports	1.38	2.12	2.04	2.43	3.34
Reading	1.16	1.45	1.66	1.96	2.26
Civic	1.57	1.92	1.77	2.06	2.24
Panel 5: ESP	75.25	71.89	71.71	71.07	71.31
Sleeping	62.78	58.68	57.78	56.94	56.74
Eating	7.33	8.13	8.58	8.75	9.49
Personal care	5.15	5.09	5.35	5.38	5.08
Total market-, nonmarket-work	46.53	53.88	55.50	57.45	57.00
Total market-, nonmarket-work + care	53.06	60.39	62.73	64.64	64.56
Leisure, sleeping	100.16	92.27	89.78	87.21	86.87
Leisure, esp	112.64	105.48	103.72	101.34	101.44
Panel 6: Other	2.31	2.13	1.55	2.01	1.99
Education	1.27	0.97	0.48	0.93	0.79
Not classified	1.03	1.16	1.07	1.08	1.20
Panel 7: Travelling	7.83	8.87	9.50	9.96	10.36
Underlying sample size	4663	4241	4344	2873	3708

* Based on data from the American Time Use Survey (ATUS) 2005-2006-2007 (pooled) (see Olmos (2017) for a detailed discussion about time allocation by income groups).

A.6 Details of the specification of the demand system

A.6.1 The EASI demand model

One of the most commonly used is to specify the cost function, $c(p, u)$, and then recover the Marshallian demands using Shephard's lemma and duality[60] (the Almost Ideal Demand System of Deaton and Muellbauer (1980) and its Quadratic version of Banks et al. (1997) are constructed with this approach).

Lewbel and Pendakur (2009) take a similar approach, but do not require invertibility of the expenditure function to recover the indirect utility. Instead, they narrow the set of expenditure functions one can choose from to construct the demand system, to a set of functions $c(p, u)$ for which there is a cardinalization of the utility level u that equals an affine transformation of the log Stone Index deflated expenditures $\log(x) - \sum_{j=1}^{J} \mathrm{w}_j \log(p_j)$, where $\log(x)$ denotes total expenditures, w_j the budget share allocated to product j, and $\log(p_j)$ its corresponding log price.

This cardinalization makes the intersection point between Hicksian demands and implicit Marshallian demands, $u = y$; thus avoiding the need of recovering the indirect utility function while leaving the functional form of $m(\cdot)$ is completely unrestricted in their dependence on implicit utility y (and observable preference shifters if added) – at the cost of having the Marshallian demands implicitly defined.[61]

The intuition behind the model is as follows (a detailed description can be found in the online appendix of Lewbel and Pendakur (2009)). By specifying the expenditure function $c(p, u)$ to take the functional form from a particular class of functions, even if we cannot invert this function to obtain indirect utility using $c(p, \psi(p, x)) = x$ (i.e. even if there is no closed form solution for indirect utility $\psi(\cdot, \cdot)$), there will be a known function $G(\cdot, \cdot, \cdot)$ depending on both p and $c(p, u)$ (but not necessarily only p and $c(p, u)$, relaxing the need of obtaining $\psi(\cdot, \cdot)$ such that $u = \psi(p, c(p, u))$) together with $\omega(p, u)$ such that

$$u = G(p, c(p, u), \omega(p, u)).$$

This identity implicitly defines u. We may define implicit utility y by $y = G(p, x, \mathrm{w})$. Then

[60] Namely, the identity, $h(p, u) = \frac{\partial c(p,u)}{\partial p}$ together with

$$g(p, x) = h(p, \psi(p, x)),$$

where $\psi(p, x)$ is obtained from inverting $c(p, \psi(p, x)) = x$.

[61] This implies endogeneity by construction, however, it can be easily dealt with and Lewbel and Pendakur (2009) find little empirical difference between the coefficients of the model in which endogeneity is accounted for and that in which it is not.

by construction $u = y$ and y is defined by observables through G.[62]

Shephard's lemma implies the Hicksian demand system $\mathrm{w} = \omega(p, u)$, where w denote observed budget shares – which depend on expenditure x and prices p,– and $\omega(p, u)$ denote Hicksian budget shares $\omega(p, u) = \left(\frac{p_1 h_1(p,u)}{c(p,u)}, \ldots, \frac{p_J h_J(p,u)}{c(p,u)}\right)$ – that depend on utility u and prices. Since $g(p, x) = h(p, u)$, then we get the implicit Marshallian demand system

$$\mathrm{w} = \omega(p, x).$$

This approach is particularly useful because the functional form of $m(\cdot)$ is completely unrestricted in their dependence on implicit utility y (and observable preference shifters if added). Hence, Engel curves may have any shape and any degree of variety across goods. In particular, it fits the tails of complex Engel curves better than the flexible QUAIDS model which are of particular interest for the paper. Additionally, nothing about the shape of Engel curves need to be known in advance and both observed and unobserved preference shifters can be easily introduced into the model. Finally, the functional form of budget shares is completely unrestricted and Gorman's rank three limit does not apply.

A.6.2 Specification of the Model

Recall, the structural model in (3) is given by

$$\mathbf{w}_{it} = \sum\nolimits_{r=1}^{R} \boldsymbol{\beta}_r y_{it}^r + \mathbf{A} \log \mathbf{p_{it}} + \boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_{it}.$$

I let the degrees of the polynomial in expenditure, R, depend on category j – thus I denote it by R_j,– as follows. Let $\Theta_j = \sum_{k=1}^{J} a_{jk} \log(p_{ikt}) + \varepsilon_{ijt}$. Then, we can rewrite the model as

$$\mathrm{w}_j = \sum_{r=1}^{R_j} \beta_{rj} y^r + \Theta_j. \tag{7}$$

I consider the correlation of w_j with y and its exponents, conditional on Θ_j, by comparing the adjusted R^2's of the regression of w_j on Θ_j, with that of w_j on y and Θ_j, of w_j on y, y^2 and Θ_j, y, y^2, y^3 and Θ_j and so on.[63] Table 5 below shows summary statistics of the R^2's corresponding to model (7), varying R_j from 0 – no income effect – to 4 – a quartic term in income. Table 6 shows summary statistics of the ratio between these R^2's. We see how the

[62]It is the dependence of $y = G(p, x, \mathrm{w})$ on w what make the resulting demand system implicit. The reason being that the demand equations will have w_j on the left and right hand side.

[63]Recall, the adjusted R^2 is adjusted for the number of predictors in the model. The adjusted R^2 increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

fit of the model – adjusting for the number of predictors in the model – improves by almost 30% on average when introducing a linear income term. The fit improves on average by 3% when introducing a quadratic term of income and by half a percentage point when adding a cubic term.

Table 5: *Adjusted* R^2

R^2 of model (7) when	Mean	Std. Dev.	Min	Max
$R_j = 0$	0.00659	0.0066007	0.0002295	0.023457
$R_j = 1$	0.0081043	0.0081393	0.0002363	0.0281206
$R_j = 2$	0.0083331	0.008436	0.0002449	0.0307689
$R_j = 3$	0.008379	0.0085083	0.0002453	0.0313763
$R_j = 4$	0.0084074	0.0085502	0.000245	0.0317141

Table 6: *Percent increase in adjusted* R^2

Ratio of R^2 in (7)	Mean	Std. Dev.	Min	Max
$R_j = 1$ with respect to $R_j = 0$	1.287574	0.3603229	1.00138	2.452227
$R_j = 2$ with respect to $R_j = 1$	1.034687	0.0635564	0.9994741	1.343368
$R_j = 3$ with respect to $R_j = 2$	1.005887	0.0116591	0.9995941	1.062162
$R_j = 4$ with respect to $R_j = 3$	1.003648	0.0071672	0.9989591	1.038382

A.7 Price Indices

A.7.1 Price imputation

3.06% of the prices paid by households are missing. I impute missing prices by regressing the average price of a barcode on market, brand, and quarter indicators; the interactions between between brand and quarter indicators, and between quarter and market[64] indicators; an indicator of whether the product is organic, whether the package includes a claim of organically grown or organic ingredients and an indicator of income of the household.

The regression was separately estimated for each food category. The predicted prices were then used to replace missing prices in the construction of price indices.

[64]By *markets* I refer to the local markets defined by Nielsen and described in Appendix A.2.1.

A.7.2 Details of the Price indices

In the Homescan Nielsen Data, prices are given at the barcode level. I define P_{ijt} as

$$P_{ijt} = \sum_{u \in U_{jt}} \frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}} P_{it}(u)$$

with U_{jt} the set of barcodes u in category j and $\frac{v_{ut}}{\sum_{u' \in U_{jt}} v_{u't}}$ defined as the (nation-wide) expenditure share on barcode $u \in U_{jt}$ with respect to the rest of the barcodes in that category at time t.[65]

The resulting price indices have a mean close to that of the average prices paid and keep significant dispersion – where the prices paid recorded in the data were considered outliers (and trimmed) if they were below the bottom 10^{th} percentile or above the 90^{th} percentile. Table 7 displays the summary statistics of both actual prices paid and the price indices for some of the products in the data.

Table 7: *Summary statistics of the price indices with respect to the average prices paid.*

	Av. Price index	Av. Prices paid
Processed meat	5.030363	5.18645
	(1.773647)	(2.882033)
Fresh meat and seafood	4.435771	4.442482
	(2.17292)	(3.442447)
Unprocessed grains	3.421305	3.404945
	(2.501273)	(3.66276)
Processed carbs	4.042108	3.818387
	(0.9294422)	(2.289862)
Sweetened beverages	1.002321	1.070387
	(0.3034897)	(0.8460429)
Fresh produce	3.291158	3.339099
	(1.303104)	(2.537162)
Dairy	3.956499	3.927383
	(1.150863)	(2.527361)
Prepared/Frozen/Canned food	3.537863	3.793739
	(0.9906852)	(2.539038)
Sweeteners, Spreads and Jams	2.77917	2.897417
	(1.622848)	(2.466761)
Butter/Margarine/Oils	3.586462	3.722341
	(1.920648)	(2.611168)

[65] 3.06% of the prices paid by households are missing. I impute missing prices using a regression of the national average price for a barcode on indicators (and interactions) of market, quarter, characteristics of the barcode and income of the household as described in Section A.7.1 of the Appendix. The predicted prices were then used to replace missing prices in the construction of price indices.

Table 8: Average Standard Deviations within and across households.

	Price Indices		Unit Values	
	within	across	within	across
Processed meat	1.189	1.774	2.602	2.882
Fresh meat and seafood	1.578	2.173	2.946	3.442
Unprocessed grains	1.617	2.501	3.014	3.663
Processed carbs	0.560	0.929	2.167	2.290
Sweetened beverages	0.196	0.303	0.737	0.846
Fresh produce	0.880	1.303	2.177	2.537
Dairy	0.788	1.151	2.362	2.527
Prepared/Frozen/Canned food	0.628	0.991	2.360	2.539
Sweeteners, Spreads and Jams	0.934	1.623	2.044	2.467
Butter/Margarine/Oils	1.145	1.921	2.128	2.611

A.8 Censoring

A.8.1 Details about the Positive sample

The Nielsen Homescan data is an unbalanced panel in which respondents' mean duration is on average 6 years (the minimum length of time they stay in the sample is 1 year and the maximum 11 – which is the total length of the panel.) Most households – with the exception of 51 out of a total of 121,900 – consume at least one of the products I consider for the analysis in each period in which they belong to the sample.

I restrict the sample to that of household-half year period pairs (i, t) in which at least one of the products I consider is consumed. The remaining 121,849 households in the data, form 840,075 household-period pairs which I will refer to observations. Only 53% of these observations correspond to a bundle of the J goods in which all of the expenditure shares are positive. Let the set of observations with this property be the *positive sample*. Details about the positive sample can be found in section (A.8.1) of the Appendix.

Households in the positive sample are observed for 2.5 years on average.

Table (9) summarizes the duration of panelists in the whole dataset and in the positive subsample, the proportion of households who do not consume one of the products at any point while they are in the sample, and the proportion of households who belong to the positive sample at a point in time, but switch into a non-consumption regime.

Figure (27) plots the share of households who consume all the goods in each semester.

Table 9: Average duration of panelists in the sample.

	Nielsen Data	Positive sample
Total number of households	121,849	61,154
Average duration of respondents	6 years	2.5 years
Proportion of households with a zero at some point	88.83%	39.02%

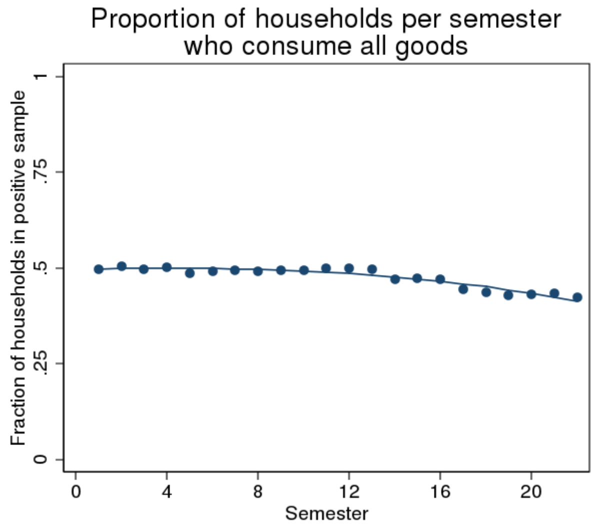


Figure 27: *Proportion of households that consume all the goods.*

A.8.2 Other approaches to deal with Censoring

To deal with the significant presence of zero expenditure on some goods I use *indirect inference* ((Gourieroux et al., 1993)).

In a typical consumer optimization problem, some households may find it optimal, under the prices they face, to consume negative amounts of a good. That is, if the non-negativity constraints for the optimization problem were disregarded, their demands for some goods would be negative. Lee and Pitt (1986) refer to these "demands" as *notional demands*. These are not true demands but rather latent variables that correspond to the observed demands via *virtual prices*, that is, the reservation prices that would induce a consumer to have zero consumption of a good. A concept introduced by Neary and Roberts (1980) to characterize household consumption behavior under quantity constraints.[66]

They use the individual's Kuhn–Tucker conditions to construct the virtual prices that would induce households to consume the exact ration levels of rationed goods had they been

[66]The paper is motivated by a context in which households are free to purchase in some markets but are forced to purchase certain levels of goods in other markets – consumption restrictions during war time are a good example of this setting.

unrationed.

Lee and Pitt (1986) restrict their attention to the case where the rationing point is zero, which greatly simplifies the virtual prices functions. Virtual prices in this setting are the prices that would induce households to satisfy the non-negativity constraints of the cost minimization problem – without imposing the non-negativity constraints.[67] They compare the virtual prices implied by corner solutions to the actual prices to derive the likelihood of observing the households' actual chosen bundle of goods.

The idea can be illustrated for the case of two goods. Suppose that the utility maximizing bundle is a corner solution where $x_1 = 0$ at market prices (p_1, p_2) as shown in Figure (28). If the consumer's optimization problem was solved without regard of the non-negativity

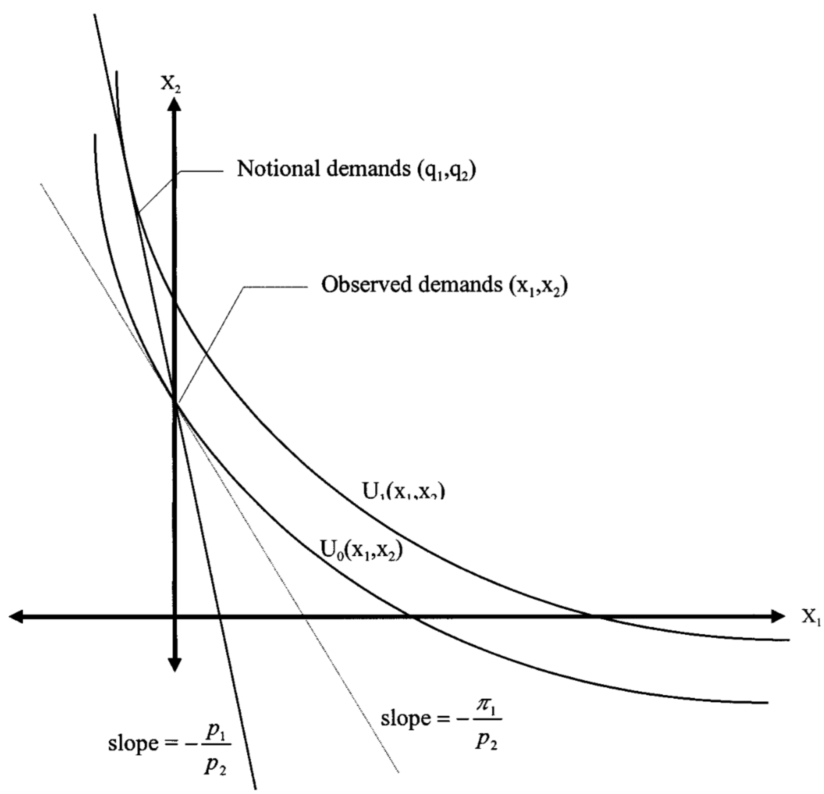


Figure 28: *Figure taken from Phaneuf (1999).*

constraint $(x_1, x_2) \geq 0$, the solution would be given by the notional demands (q_1, q_2), where $q_1 < 0$.

The virtual price π_1 for good 1 is a reservation price for which consumption would be exactly 0. By using the price ratio $\frac{\pi_1}{p_2}$ rather than $\frac{p_1}{p_2}$, one can construct a tangency condition

[67]This is the dual of the approach developed by Wales and Woodland (1983), who constructed the likelihood function based on the Kuhn-Tucker conditions of maximization of a stochastic direct utility function subject to budget and non-negativity constraints.

for the observed consumption bundle that can be used to form an estimating equation. Since market prices are below the virtual prices of non-consumed goods, the comparison among these can be used to construct a likelihood function to estimate the parameters of the model using all observed bundles.

These virtual prices are themselves functions of the household's income and of the prices of the uncensored goods. They can be calculated from the unconstrained demand and supply functions as I illustrate in section A.8.3.

This approach becomes computationally infeasible as the number of censored goods grows. Because of this, many studies using highly disaggregated data like the Nielsen Homescan Data, apply the Tobit estimator to estimate latent demand (see for example Zhen et al. (2013)). This approach circumvents the empirical difficulties of imposing non-negativity restrictions under the framework outlined above. However, for the case of corner solutions in demand system estimation, the application of Tobit estimation will, for systems with more than two goods, result in biased estimates since they fail to consider that consumers response to price depends on the set of goods it consumes at corners.

Furthermore, excluding from the sample those observations in which kink points are observed is likely to result in sample selection bias (a widely used specification of preferences with highly disaggregated data such as the Nielsen Data comes from the CES family[68]).

To deal with the significant presence of zero expenditure on some goods without resorting to the computationally intensive approach developed by Lee and Pitt (1986), I use a simulated method of moments estimator.

A.8.3 The use of virtual prices in my setting

In this case, the structural demand model consists of a system of equations given by (1) which I repeat below for convenience.

$$\mathrm{w}_{ijt} = \alpha_j + \sum_{r=1}^{R} \beta_{rj} y_{it}^{r} + \sum_{k=1}^{J} a_{jk} \log(p_{ikt}) + \xi_i + \varepsilon_{ijt}, j \in \{1, \ldots, J\}.$$

Given a consumption regime, for example $\{w_{i1}, \ldots, w_{iJ}\}$, where the first ℓ goods are not consumed, we can solve for the virtual prices that support the zero demands for goods $1, \ldots, \ell$ by substituting zero for $w_{i1}, \ldots, w_{i\ell}$ in the demand system (1) and solving for the reservation prices $\pi_{i1}, \ldots, \pi_{i\ell}$ that support this regime

[68] Dubois et al. (2014), Handbury (2013) and Jaravel (2016) for example make this functional form assumption.

$$\begin{bmatrix} \log(\pi_1) \\ \log(\pi_2) \\ \vdots \\ \log(\pi_\ell) \end{bmatrix} = -A_{1,\ell}^{-1} \begin{bmatrix} \alpha_1 + \sum_{r=1}^{R} \beta_{r1} y_{it}^r + \sum_{k=\ell+1}^{J} a_{1k} \log(p_{ikt}) + \xi_i + \varepsilon_{i1t} \\ \vdots \\ \alpha_\ell + \sum_{r=1}^{R} \beta_{r\ell} y_{it}^r + \sum_{k=\ell+1}^{J} a_{\ell k} \log(p_{ikt}) + \xi_i + \varepsilon_{i\ell t} \end{bmatrix}$$

where

$$A_{1,\ell} = \begin{bmatrix} a_{11} & \dots & a_{1\ell} \\ \vdots & \ddots & \vdots \\ a_{\ell 1} & \dots & a_{\ell\ell} \end{bmatrix}$$

given these virtual prices and a specification of the joint density f_ε for ε, we can define the likelihood function for estimation.

There are 2^{J-1} possible consumption regimes for which the likelihood function for estimation, which is formed by the product of the appropriate probabilities, can be constructed. Maximum likelihood is used to recover estimates of the parameters. This is not an attractive option when J is large. For example, when J is 10, 2^{J-1} is 512, and high dimensional integrals are involved for many of the combinations. For example, in cases where ℓ is 7, one must evaluate a 7 dimensional integral. This is why I use a simulation based estimator.

A.8.4 Bias due to censoring

I find the bias due to censoring to be small under the fixed effect specification I use.[69] To simplify the following discussion, let me separate the goods into healthy and unhealthy and define the positive sample as the set of (i, t) observations for which consumption of both "goods" is positive.

Suppose we were to restrict the sample to that of households who consume positive amounts of the healthy and unhealthy categories. Estimates of the structural parameters in this sample would yield unbiased estimates if the selection into consuming positive amounts was a deterministic function of the covariates[70]

If there was an unobserved factor that drives the decision of a household to consume positive amounts of a good, and this unobserved factor was correlated with the random component of the demand system, then these estimates would be biased.

[69] As can be seen in the discussion below, in a cross-sectional analysis, the bias due to censoring could be sizable.

[70] Or, a function of the covariates with a random component that is uncorrelated with the random heterogeneity of the demand system.

For example, let the distribution of the fixed effects ξ_{ij}[71] in the population be fixed and suppose that a higher stock of preferences ξ_{ij} increases the returns to consumption of healthful products, so that ξ_{ij} is positively correlated with the expenditure share on the healthy category.

Suppose, additionally, that expenditure is correlated with positive consumption of healthful goods. Then upon an increase in the budget, the proportion of households who consume zero of the healthful category would decrease. However, a larger proportion of households who switch from zero consumption to positive consumption of healthful products would be those with lower ξ_{ij} (those with high ξ_{ij} were consuming healthful products already). Then the higher income households who consume a positive amount of the healthful category have, on average, lower ξ_{ij} than the population.

Consider restricting the sample to those households who consume a positive amount of both the healthful and the unhealthful categories. On the one hand, the increase in income will induce some households who had zero consumption of the healthy category to enter the sample, lowering the average ξ_{ij} of the sample with respect to the population level. On the other hand, households with high ξ_{ij} that were in the sample may shift all their consumption towards the healthy category, and drop out of the sample, lowering further the average human capital of the sample.

This selection process, thus, causes a negative correlation between income and ξ_{ij}, causing a downward bias in the coefficient on income in the demand equations.

This would be a serious concern in the cross-section because both income and prices could be indeed correlated with positive consumption of food products in a non-trivial way through ξ_{ij}.

In the time series, however, this problem is attenuated. Recall, each demand equation of the demand system reads

$$\mathrm{w}_{ijt} = \alpha_j + \beta_j y_{it} + \sum_{k=1}^{J} a_{jk} \log(p_{ikt}) + \xi_{ij} + \varepsilon_{ijt}.$$

Let the selection rule that assigns individuals in the overall sample population to the restricted sample be given by

$$d_{it} = \mathbb{1}(z_{it}\gamma_j + u_{ijt}).$$

[71]Let us think of ξ_{ij} as a collection of traits such as knowledge or nutritional awareness possessed by the members of the households.

If $f(\varepsilon_1, \ldots, \varepsilon_J) = f_1(\varepsilon_1) \cdots f_J(\varepsilon_J)$, then,[72] differencing the demand equations over time or using the within transformation, would eliminate any potential selection problem which operates through the fixed effect ξ_{ij}. Indeed, a sufficient condition for identification taking differences $\mathrm{w}_{ijt} - \mathrm{w}_{ijs}$ for $t \neq s$ and assuming that the unobserved disturbances are independent is

$$\mathbb{E}(\varepsilon_{ijt} - \varepsilon_{ijs} | X_{ijt}, X_{ijs}, d_{it} = d_{is} = 1) = 0. \tag{8}$$

[73]

Condition (8) puts no restrictions on how the selection mechanism or the regressors relate to ξ_{ij}. That is, as long as the selection process is uncorrelated with changes in the unobserved disturbances and households who are selected in a time period, are also selected in the others, the difference estimator will lead to unbiased estimates of the structural parameters.

The threat to identification comes from time-varying unobserved disturbances affecting selection in and out of the sample. In the example above, this threat would materialize if human capital changed over time in a way that would induce households to be selected into the sample in some periods and selected out in others.[74]

A.9 Markov Chain Monte Carlo Simulation

The Indirect Inference objective function is given by

$$\hat{\boldsymbol{\gamma}} = \text{argmin}_{\gamma} \left(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}(\tilde{\mathbf{w}}_{it}(\boldsymbol{\gamma})) \right)' \Omega \left(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}(\tilde{\mathbf{w}}_{it}(\boldsymbol{\gamma})) \right). \tag{9}$$

where the weight matrix Ω is a diagonal matrix using the inverse variance of each moment in the data.

Chernozhukov and Hong (2003) propose a Markov-Chain Monte-Carlo estimation procedure for this problem. MCMC is a derivative-free method that circumvents the curse of dimensionality because it only requires evaluating the objective function at many different points to simulate a chain of parameters that converges to a probability distribution over

[72] Estimating (1) in levels by pooled OLS (or in the cross-section using one time-period) would require the following condition to be satisfied for the estimated parameters to be consistent.

$$\mathbb{E}(\xi_{ij} + \varepsilon_{ijt} | X_{ijt}, d_{it} = 1) = 0.$$

[73] See Dustmann and Rochina-Barrachina (2007).

[74] Note that if households are selected at a given point in time t_0 and stay in the sample continuously until another point in time T_0 and not observed in any other period in the selected sample, there would be no threat to consistency of the estimates either (in the example above this would mean that changes in human capital are "smooth" or "monotonic.")

the parameter vector,

$$p(\theta) = \frac{e^{L_n(\theta)}\pi(\theta)}{\int e^{L_n(\theta)}\pi(\theta)d\theta}$$

The estimator is the average over the K elements of the converged chain

$$\hat{\theta} = \frac{1}{K}\sum_{k=1}^{K}\theta^{(k)}.$$

In practice, I use the Metropolis-Hastings algorithm. One starts from a parameter guess $\theta^{(k)}$ and generates an alternative draw θ' from a proposal density $q(\theta'|\theta^{(k)})$ which I assume to be a random walk with multivariate normal distribution. I update the parameter guess according to

$$\theta^{(k+1)} = \begin{cases} \theta' & \text{with probability} \quad \rho(\theta^{(k)}, \theta') \\ \theta^{(k)} & \text{with probability} \quad 1 - \rho(\theta^{(k)}, \theta') \end{cases}$$

where $\rho(x, y) = \min\left(e^{L_n(y)-L_n(x)}, 1\right)$ under the assumption of uniform priors and the proposal density a random walk.

A.10 Estimation Results

A.10.1 Real expenditure and Income

Nominal expenditures $\log(x)$ are increasing in income.

However, the Stone index deflated log nominal expenditures, $\log(x) - \mathbf{w}\log(\mathbf{p})$ are not as can be seen in figure 30.

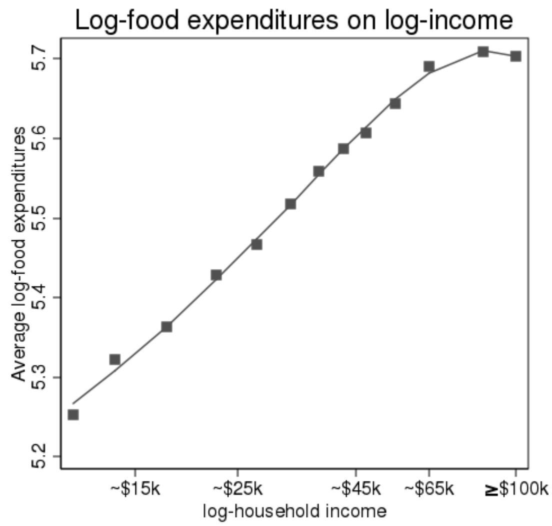


Figure 29: *Average log of expenditure on food by income level.*

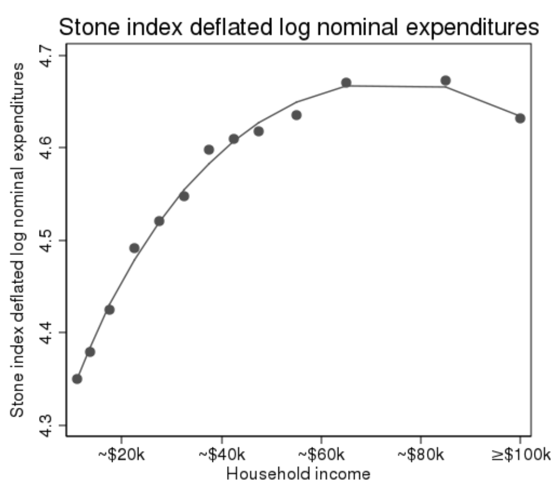


Figure 30: *Average log of expenditure on food by income level.*

A.10.2 Estimated Parameters

Table 10 reports the estimates of the coefficient on the expenditure term – each parameter estimate represents 0.01 times the change upon a 1% increase in expenditure.

The first column corresponds to the estimated parameters from the auxiliary model defined on the positive sample; and the second column corresponds to the structural parameters in model (3).

Table 10: *Estimation Results: Income effects*

	Estimated Coefficients	
	Positive Sample	Whole Sample
Processed meat	0.0175	0.0138***
	(0.0168)	(0.0036)
Fresh meat and seafood	0.1854***	0.1293***
	(0.0318)	(0.0048)
Unprocessed grains	−0.0133	−0.0147***
	(0.0090)	(0.0018)
Processed carbs	−0.0576*	−0.0091
	(0.0242)	(0.0053)
Sweetened beverages	−0.0170	−0.0432***
	(0.0200)	(0.0059)
Fruit and Vegetables	0.0879***	0.0763***
	(0.0273)	(0.0057)
Dairy	−0.0880***	−0.0387***
	(0.0239)	(0.0040)
Prepared/Frozen/Canned food	−0.0951***	−0.0926***
	(0.0264)	(0.0070)
Sweeteners, Spreads and Jams	−0.0031	−0.0049**
	(0.0097)	(0.0020)
Butter/Margarine/Oils	−0.0167**	−0.0265***
	(0.0065)	(0.0015)

* **Note:** The reported standard errors of the coefficients of the whole sample are based on 50 bootstrap replications of the estimation procedure.

The price effects captured by the parameters a_{jk} are compensated budget price semielasticities. That is, a_{jk} captures the effect of a change in log-prices on expenditure shares, while compensating the agent to keep him in the same indifference curve.

Table 11a and Table 11b report the estimated price effects from the auxiliary model defined on the positive sample; and Table 12a and Table 12b the corresponding estimates of the structural price effects on expenditure shares.

When so many cross-price elasticities are estimated with minimal functional form restrictions, some elasticities will not be easy to interpret. However, for the few that I have a priori expectations, the estimated cross price elasticities align: in particular, one would expect that fruit and vegetables and fresh meat and seafood are complements as they are both typically used for preparing a meal. Similarly, dairy and sweeteners spreads and jams contain common ingredients that are consumed together (for example as a breakfast).

***Table 11:** Estimation Results*

***(a)** Price effects from the positive sample: product categories 1 to 5*

		Product Number				
		1	2	3	4	5
1	Processed meat	0.049*** (0.003)	−0.017*** (0.005)	−0.002 (0.001)	−0.008* (0.004)	−0.011** (0.003)
2	Fresh meat and seafood	−0.004 (0.004)	0.037*** (0.008)	−0.002 (0.002)	−0.014* (0.006)	−0.018*** (0.005)
3	Unprocessed grains	−0.002*** (0.001)	−0.004*** (0.001)	0.012*** (0.000)	−0.003*** (0.001)	−0.001 (0.001)
4	Processed carbs	−0.010 (0.005)	−0.015 (0.010)	−0.008* (0.003)	0.068*** (0.008)	−0.008 (0.006)
5	Sweetened beverages	−0.004 (0.004)	−0.042*** (0.008)	0.000 (0.002)	0.015* (0.006)	0.035*** (0.005)
6	Fruit and Vegetables	−0.004** (0.002)	0.022*** (0.003)	−0.003*** (0.001)	−0.013*** (0.002)	−0.009*** (0.002)
7	Dairy	0.011*** (0.002)	−0.003 (0.003)	−0.003** (0.001)	0.001 (0.003)	−0.006** (0.002)
8	Prepared/Frozen/Canned food	−0.006* (0.003)	−0.019*** (0.005)	−0.001 (0.001)	0.006 (0.004)	−0.003 (0.003)
9	Sweeteners, Spreads and Jams	−0.003* (0.002)	−0.015*** (0.003)	0.001 (0.001)	0.001 (0.002)	−0.003 (0.002)
10	Butter/Margarine/Oils	−0.002* (0.001)	−0.011*** (0.002)	0.001* (0.001)	−0.002 (0.001)	−0.001 (0.001)

***(b)** Price effects from the positive sample: product categories 6 to 10*

		Product Number				
		6	7	8	9	10
1	Processed meat	−0.012** (0.004)	0.011** (0.004)	−0.003 (0.004)	−0.005*** (0.002)	−0.003* (0.001)
2	Fresh meat and seafood	0.023*** (0.007)	−0.005 (0.006)	−0.008 (0.006)	−0.008*** (0.002)	−0.002 (0.002)
3	Unprocessed grains	−0.002** (0.001)	0.000 (0.001)	−0.001 (0.001)	0.000 (0.000)	0.000 (0.000)
4	Processed carbs	0.015 (0.009)	−0.014 (0.008)	−0.010 (0.008)	−0.014*** (0.003)	−0.005* (0.002)
5	Sweetened beverages	−0.014* (0.007)	0.003 (0.006)	0.010 (0.006)	−0.004 (0.002)	0.001 (0.002)
6	Fruit and Vegetables	0.047*** (0.002)	−0.011*** (0.002)	−0.018*** (0.003)	−0.006*** (0.001)	−0.004*** (0.001)
7	Dairy	−0.008** (0.003)	0.008** (0.003)	−0.001 (0.003)	0.000 (0.001)	0.002** (0.001)
8	Prepared/Frozen/Canned food	−0.013** (0.004)	0.014*** (0.004)	0.018*** (0.004)	0.003 (0.001)	0.001 (0.001)
9	Sweeteners, Spreads and Jams	−0.009*** (0.003)	0.004 (0.002)	0.003 (0.002)	0.019*** (0.001)	0.001* (0.001)
10	Butter/Margarine/Oils	−0.005*** (0.002)	0.005*** (0.001)	0.002 (0.002)	0.002*** (0.001)	0.012*** (0.000)

A.10.3 Robustness Check

As discussed in Section 5, following Banks et al. (1997), I use log-income as an instrument for expenditure. The validity of this instrument, lies in the assumption that households

Table 12: Estimation Results

(a) Price effects from the whole sample: product categories 1 to 5

		Product Number				
		1	2	3	4	5
1	Processed meat	0.048^{***} (0.0012)	-0.011^{***} (0.0019)	-0.004^{***} (0.0005)	-0.016^{***} (0.0017)	-0.004^{***} (0.0011)
2	Fresh meat and seafood	-0.011^{***} (0.0019)	0.041^{***} (0.0020)	-0.008^{***} (0.0007)	0.010 (0.0028)	-0.033^{***} (0.0016)
3	Unprocessed grains	-0.004^{***} (0.0005)	-0.008^{***} (0.0007)	0.013^{***} (0.0004)	-0.006^{**} (0.0011)	0.006^{***} (0.0009)
4	Processed carbs	-0.016^{***} (0.0017)	0.010 (0.0028)	-0.006^{**} (0.0011)	0.045^{***} (0.0020)	0.003 (0.0016)
5	Sweetened beverages	-0.004^{***} (0.0011)	-0.033^{***} (0.0016)	0.006^{***} (0.0009)	0.003 (0.0016)	0.041^{***} (0.0015)
6	Fruit and Vegetables	-0.004^{***} (0.0009)	0.024^{***} (0.0012)	-0.007^{***} (0.0006)	-0.016^{***} (0.0013)	-0.010^{***} (0.0007)
7	Dairy	0.006^{***} (0.0009)	-0.001^{**} (0.0014)	0.002^{**} (0.0006)	-0.004^{***} (0.0010)	-0.001 (0.0011)
8	Prepared/Frozen/Canned food	-0.008^{***} (0.0011)	-0.026^{***} (0.0016)	0.000^{*} (0.0009)	0.000 (0.0013)	0.001^{*} (0.0013)
9	Sweeteners, Spreads and Jams	-0.005^{***} (0.0005)	-0.015^{***} (0.0009)	0.000^{**} (0.0004)	0.001 (0.0007)	-0.001 (0.0007)
10	Butter/Margarine/Oils	-0.002^{***} (0.0004)	-0.010^{***} (0.0007)	0.001^{***} (0.0004)	-0.004^{***} (0.0005)	0.004^{***} (0.0005)

(b) Price effects from the whole sample: product categories 6 to 10

		Product Number				
		6	7	8	9	10
1	Processed meat	-0.004^{***} (0.0009)	0.006^{***} (0.0009)	-0.008^{***} (0.0011)	-0.005^{***} (0.0005)	-0.002^{***} (0.0004)
2	Fresh meat and seafood	0.024^{***} (0.0012)	-0.001^{**} (0.0014)	-0.026^{***} (0.0016)	-0.015^{***} (0.0009)	-0.010^{***} (0.0007)
3	Unprocessed grains	-0.007^{***} (0.0006)	0.002^{**} (0.0006)	0.000^{*} (0.0009)	0.000^{**} (0.0004)	0.001^{***} (0.0004)
4	Processed carbs	-0.016^{***} (0.0013)	-0.004^{***} (0.0010)	0.000 (0.0013)	0.001 (0.0007)	-0.004^{***} (0.0005)
5	Sweetened beverages	-0.010^{***} (0.0007)	-0.001 (0.0011)	0.001^{*} (0.0013)	-0.001 (0.0007)	0.004^{*} (0.0005)
6	Fruit and Vegetables	0.043^{***} (0.0010)	-0.005^{***} (0.0014)	-0.004^{***} (0.0014)	-0.011^{***} (0.0010)	-0.001^{***} (0.0007)
7	Dairy	-0.005^{***} (0.0014)	0.004^{**} (0.0012)	0.013^{***} (0.0012)	0.001^{*} (0.0007)	0.002^{***} (0.0005)
8	Prepared/Frozen/Canned food	-0.004^{***} (0.0014)	0.013^{***} (0.0012)	0.017^{***} (0.0014)	0.000^{*} (0.0007)	-0.001^{***} (0.0006)
9	Sweeteners, Spreads and Jams	-0.011^{***} (0.0010)	0.001^{*} (0.0007)	0.000^{*} (0.0007)	0.016^{***} (0.0004)	0.002^{***} (0.0003)
10	Butter/Margarine/Oils	-0.001^{*} (0.0007)	0.002^{***} (0.0005)	-0.001 (0.0006)	0.002^{***} (0.0003)	0.012 (0.0003)

face variation in income over time, while the rest of their characteristics – in particular their consumption environment, – remains unchanged. As a check of the validity of this assumption, I use different subsamples (of the positive sample) to compare estimates of the parameters.

Note that a raise in salary for one or both heads of the households while they both remain in the same firm (and thus interact with the same network of people) and without changing home address would be the type income change that satisfies the exclusion restriction – as long as the random component of preferences is uncorrelated with income.

I consider five subsamples of the positive sample. The first subsample I consider is that of households from the positive sample before they change zipcodes. This slightly reduces the number of periods I observe households – and reduces further the positive sample I can use to estimate the model. In particular, 95.3% of the household-period observations remain in the non-mover sample.[75] The estimates are very similar to the ones I get in the whole sample except for coefficients that are not significant: they remain non-significant, but these coefficients do differ from the originals.

The second subsample I consider, is that of non-movers before the male head of the household changes occupation – if at all.[76] This, again, leaves the original sample of households intact but further reduces the household-period observations to 76.9% of the original sample.[77] The estimates are again similar to those from the original positive sample except for the coefficients that were not statistically significantly different from zero.

The third, is that of non-movers before the male head and the female head change occupation. This leaves 68.26% of the household-period observations.[78] The forth, is the subsample of non-movers before the male head changes occupation and employment hours. This leaves 75% of the household-period observations.[79] and the fifth, adds the constraint that the female head does not change occupation nor employment hours. This latter subsample would be the cleanest subsample to make a comparison with the positive sample. However, this subsample contails only 66% of the household-period observations, which significantly reduces the variation that can be used for estimation.[80]

With the third, forth and fifth sub-samples, most of the coefficients become insignificant and I find larger discrepancies. However, the conclusions from the estimates (which goods are normal, which goods are inferior) remain the same except for the sweeteners, spreads and jams category, which is statistically insignificant in the positive sample (and remains statistically insignificant in all the sub-samples I consider.)

[75] In the positive sample, households are observed for an average of 4.3 periods, with a minimum of 2 periods and a maximum of 16. In the non-movers' sample, the average number of periods decreases slightly to 4.2 periods.

[76] If there is no male head, then I use the female head's occupation and consider the household before the female head changed occupation – if at all.

[77] The average number of periods I observe the households reduces to 3.5.

[78] The average number of periods households are observed in this sample is of 3.2.

[79] The average number of periods I observe households in this sample is 3.4.

[80] And an average number of periods of 3.4 again.

Table 13: *Estimated income effects from different subsamples of the positive sample.*

	Positive Sample	Subsample I	II	III	IV	V
Processed meat	0.0175 (0.0168)	0.0106 (0.0136)	0.0109 (0.0266)	0.0582 (0.0567)	0.0067 (0.0326)	0.0478 (0.0798)
Fresh meat	0.1854*** (0.0318)	0.1535*** (0.0223)	0.1982*** (0.0514)	0.2864* (0.1390)	0.2232*** (0.0691)	0.3949 (0.2675)
Unprocessed grains	−0.0133 (0.0090)	−0.0121 (0.0075)	−0.0083 (0.0149)	−0.0320 (0.0310)	−0.0010 (0.0183)	−0.0315 (0.0441)
Processed carbs	−0.0576* (0.0242)	−0.0329 (0.0190)	−0.0685 (0.0389)	−0.0850 (0.0800)	−0.1183 (0.0543)	−0.1786 (0.1533)
Sweetened beverages	−0.0170 (0.0200)	−0.0255 (0.0166)	−0.0726* (0.0367)	−0.1568 (0.0986)	−0.0916 (0.0482)	−0.2188 (0.1772)
Fruit and Vegetables	0.0879*** (0.0273)	0.0846*** (0.0222)	0.1405** (0.0486)	0.1067 (0.0881)	0.1789** (0.0661)	0.1364 (0.1352)
Dairy	−0.0880*** (0.0239)	−0.0874*** (0.0195)	−0.0846* (0.0382)	−0.0936 (0.0767)	−0.0720 (0.0457)	−0.0533 (0.1017)
Prepared/preserved food	−0.0951*** (0.0264)	−0.0760*** (0.0202)	−0.1133** (0.0439)	−0.1051 (0.0852)	−0.1202* (0.0550)	−0.1231 (0.1284)
Spreads and Jams	−0.0031 (0.0097)	−0.0035 (0.0078)	0.0074 (0.0157)	0.0280 (0.0340)	0.0069 (0.0192)	0.0354 (0.0503)
Butter/Margarine/Oils	−0.0167** (0.0065)	−0.0113** (0.0052)	−0.0097 (0.0104)	−0.0068 (0.0204)	−0.0128 (0.0129)	−0.0090 (0.0293)

* **Note:** Each column reports the estimated coefficients on log-deflated expenditure for a different subsample. *Subsample I* refers to the subsample of non-movers. *Subsample II* refers to the subsample of non-movers before the male head of the household changes occupation. *Subsample III* refers to the subsample of non-movers before the male head and the female head change occupation. *Subsample IV* non-movers before the male head changes occupation or employment hours. *Subsample V* non-movers before the either head changes occupation or employment hours.

A.10.4 Counterfactual Quantities Consumed

The difference between the predicted shares, $\mathbb{E}(\mathrm{w}_{ijt} \mid I_{it} = I)$, and the counterfactual shares under the price schedule of the top income group $(p_k^H)_{k=1}^J$, $\mathbb{E}(\mathrm{w}_{ijt} \mid I, p_{ikt} = p_k^H)$, tell us by how much households in income group I would adjust their expenditure shares in each category if they all faced prices $p_1^H, p_2^H, \ldots, p_J^H$.

$$\mathbb{E}(\mathrm{w}_{ijt} \mid I, p_{ikt} = p_k^H) - \mathbb{E}(\mathrm{w}_{ijt} \mid I) = \sum_{k=1}^{J} \hat{a}_{jk} \left(p_k^H - p_{ikt}(I)\right). \tag{10}$$

That is, the left-hand side of (10) captures the effect on i's consumption of category j of "changing" the price schedule from $(p_{ikt}(I))_{k=1}^J$ to the average log-price schedule from the top income group $(p_{ikt}^H)_{k=1}^J$.

I normalize the response in this shift in prices by converting the counterfactual shares into the implied quantities fixing food expenditure at a hundred dollars for all households. I plot the predicted quantities for each food category by income group in Figure 31.[81]

[81]Note that the y-axes are not in the same scale.

A.11 Counterfactual Analysis

A.11.1 Price Effects

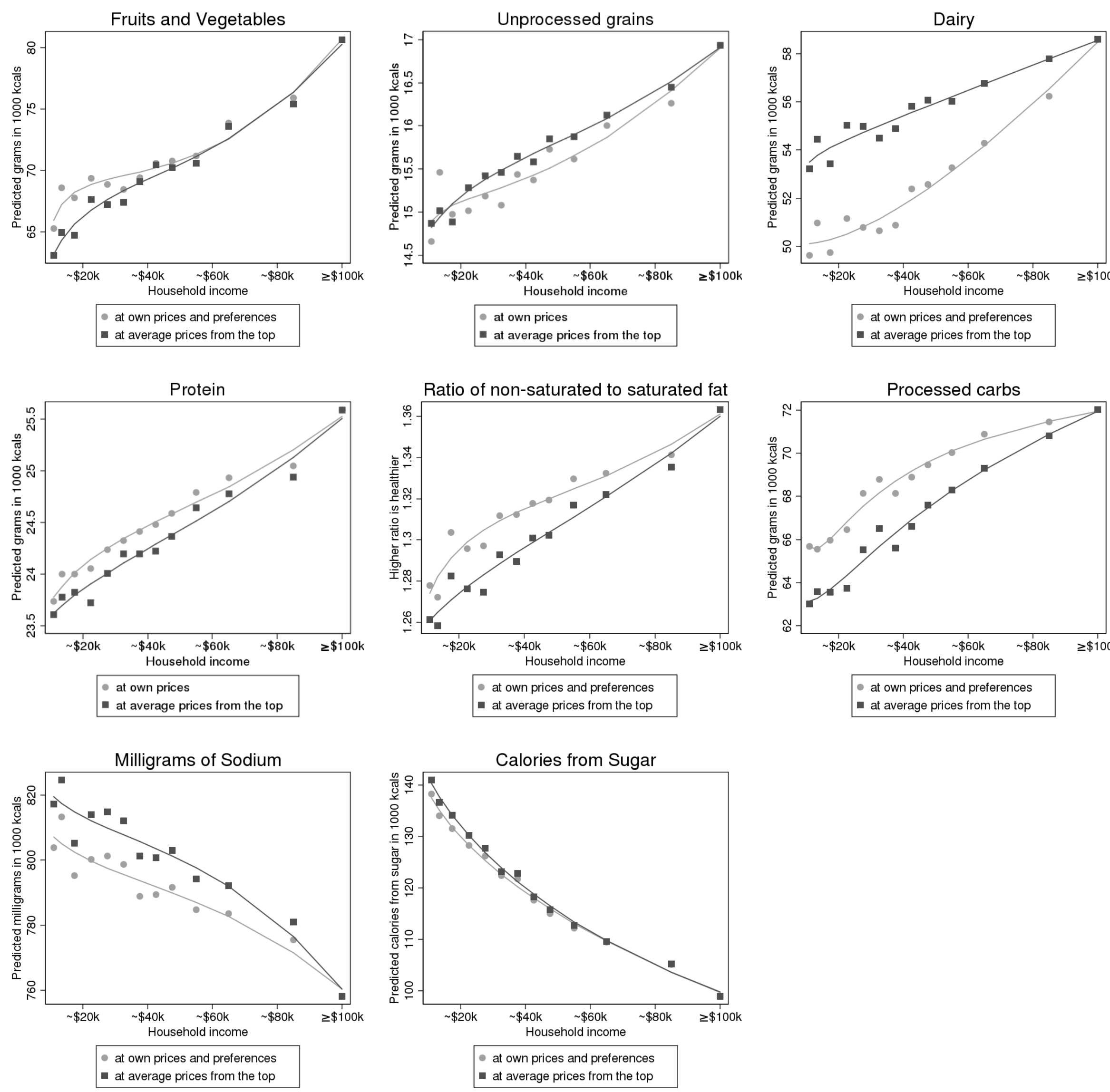


Figure 31: Price effects.

A.11.2 Preferences

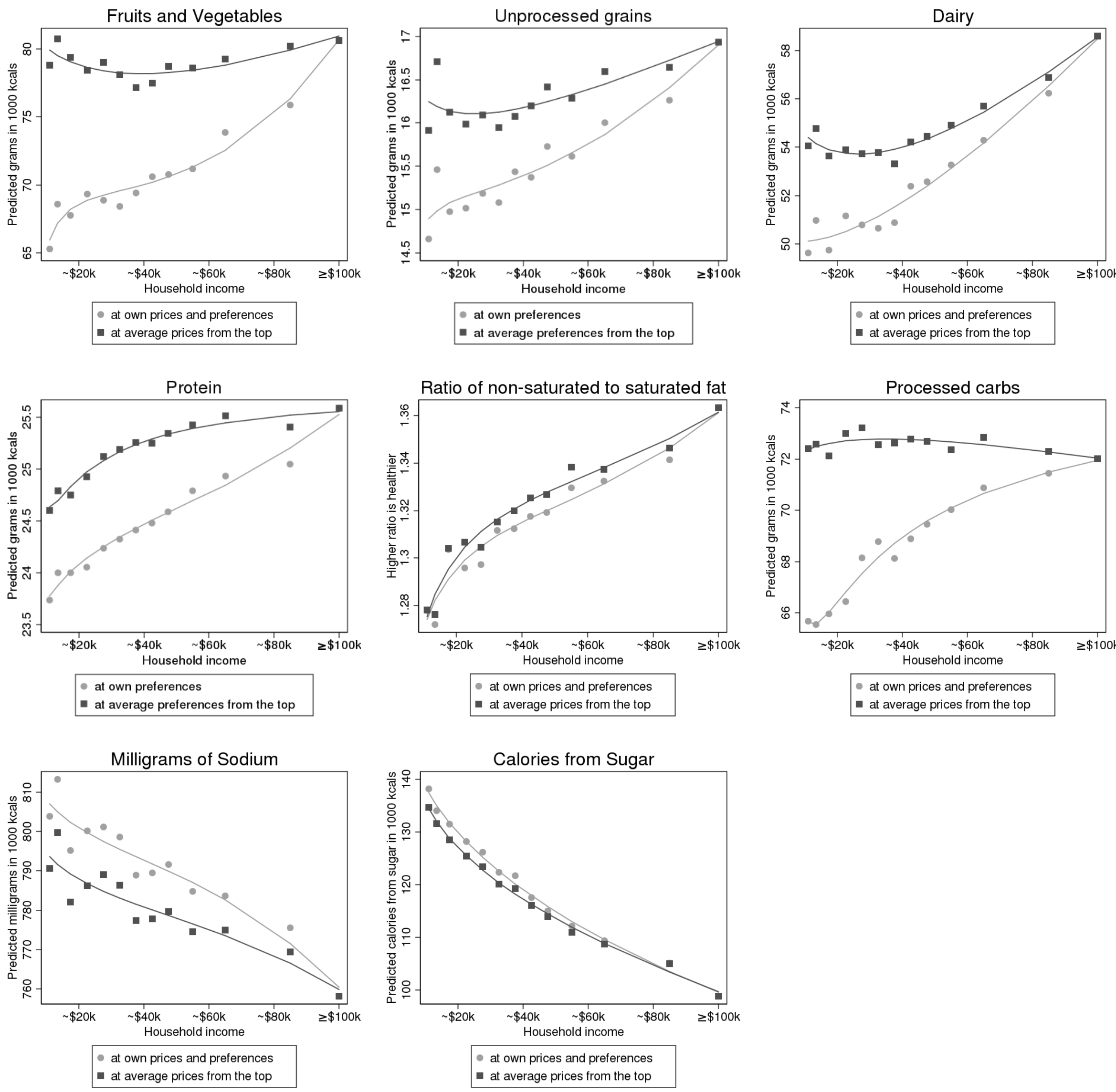


Figure 32: *Preference effects.*

A.11.3 Income Effects

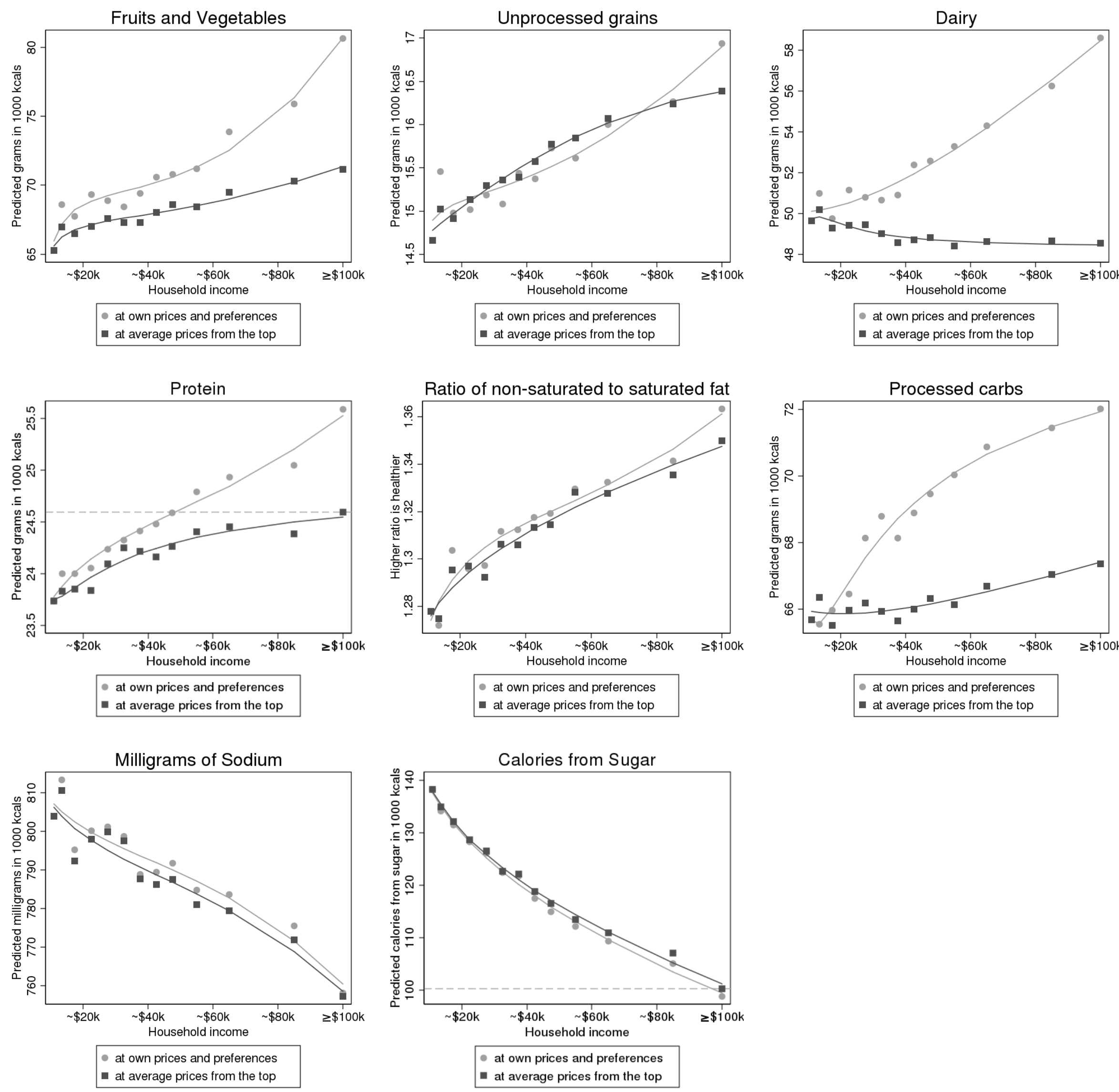


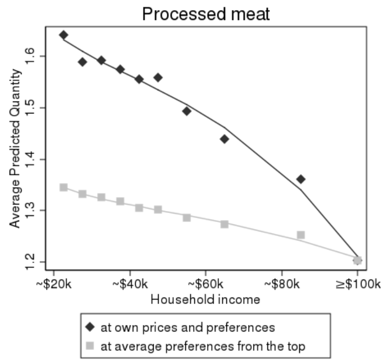
Figure 33: *Income effects.*

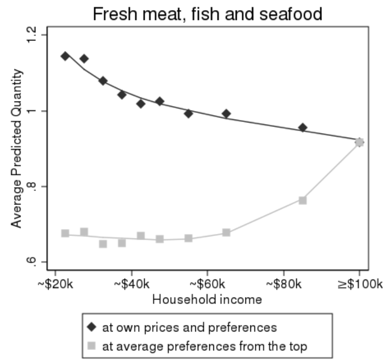
A.11.4 Robustness Check

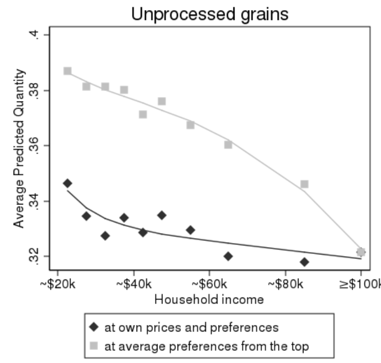
	Occupation Groups	Description
1	**Professional Specialty Occupations**	Economist, Mathematician, Scientist, Researcher Accountant, Auditor Architect, Engineer, Pilot Artist, Entertainer, Writer, Dancer, Composer, Athlete Business professional, non- managerial, Computer Programmer, System Analyst, Data Processor Dentist, Doctor, Pharmacist, Physician's Assistant, Psychologist Educator, Lecturer, Librarian, Teacher, Coach, Lawyer, Paralegal, Judge, Medical Technician, Paramedic, RN, Therapist, Social Worker Religious, Clergy Member
2	**Managerial and Public Administration**	Administrator, Company, Officer, Manager, Supervisor Banker, Controller, Financial Analyst Builder, Contractor, Landscaper, Owner of Business, Company, or Store Buyer, Purchasing Agent, Public Official, Politician, Government Employee
3	**Communications, and Professional Services**	Bank Teller, Bookkeeper, Cashier Clerk, Gas Attendant, Stock, Inventory Computer, Graphic Design, Insurance Adjuster, Underwriter Mailroom, Messenger, Postal Worker Receptionist, Secretary, Typist, Data Entry
4	**Sales**	Sales - Industrial, Wholesale Sales - Insurance, Real Estate, Services Sales - Retail
5	**Technician and Public Services**	Foreman, Baker, Butcher, Seamstress, Tailor Carpenter, Electrician, Painter, Plumber, Exterminator, Construction or Road Machine Operator Mechanic, Repairman, Technician (except medical), Utility Lineman, Serviceman, Building Inspector
6	**Transportation**	Factory Machine Operator, Delivery, Route man, Driver-Bus, Taxi, Truck, Factory, Transportation Worker (Airline, Railroad, Cruise)
7	**Military**	Member of Armed Forces
8	**Service Occupations**	Barber, Beautician, Nail Technician, Salon, Makeup Artist Bartender, Chef, Food Service, Hotel Service, Child Care Worker, Housekeep, Maid, Dental Assistant, Practical Nurse, Dental Hygienist, Fire fighter, Police Officer, Sanitation, Security Officer Janitor, Porter
9	**Agriculture**	Farmer (Manager, Owner, Worker)
10	**Student**	Students Employed $<$30 hours
11	**Operators, Fabricators, and Laborers**	Construction worker, Shipping and Receiving Fisherman, Gardener, Lumberman
12	**Non-employment**	Housewife, Retired, Unable to work Unemployed, Laid off

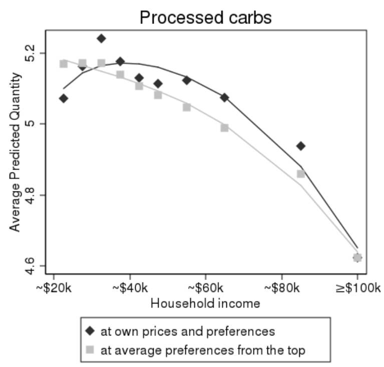
The occupation groups are not named in the annual questionnaire given to the respondents. Households choose their "occupation group" based on the list of occupations displayed in the left panel.

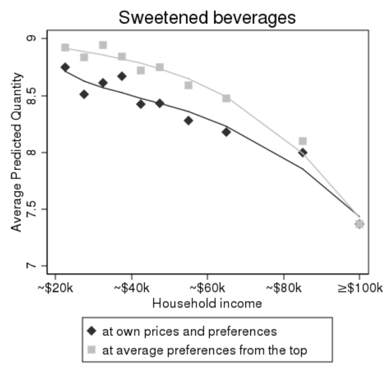
A.11.5 Preferences

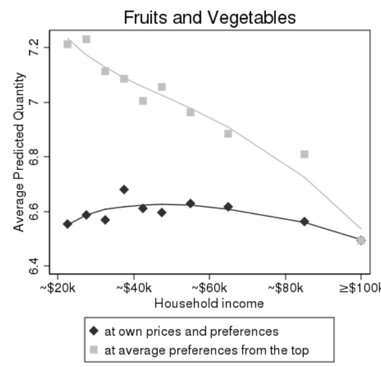


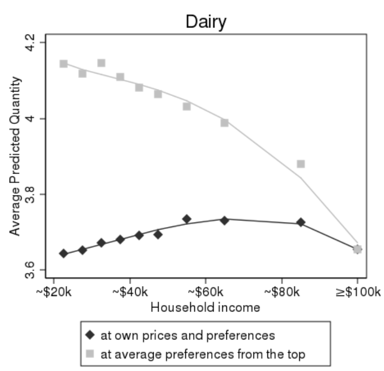


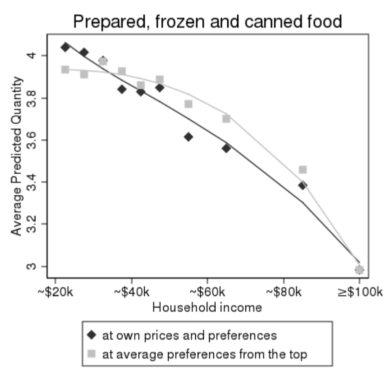


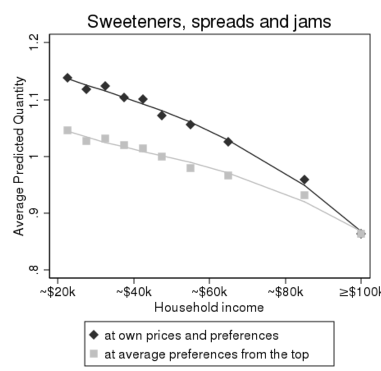


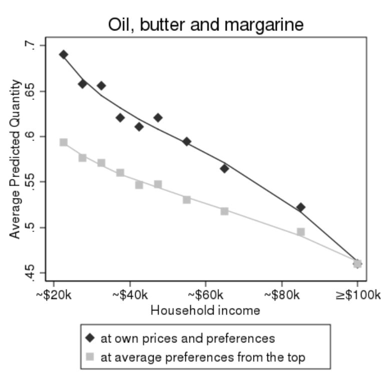




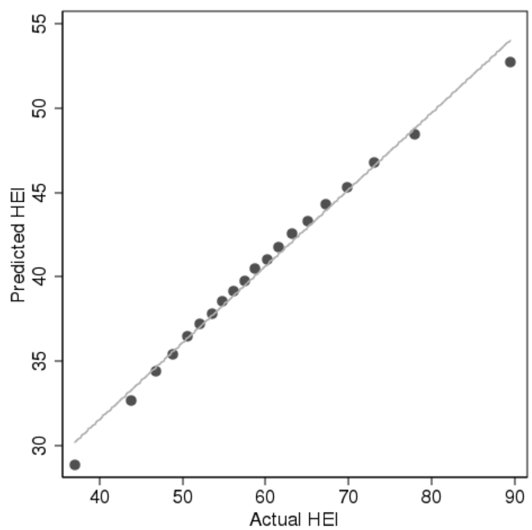


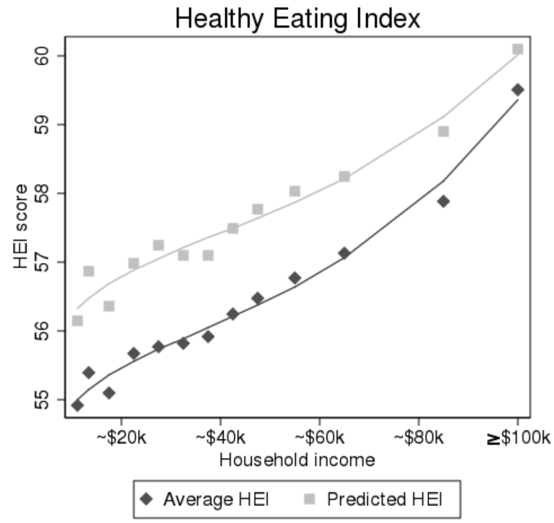






A.11.6 Counterfactual Nutrient Consumption

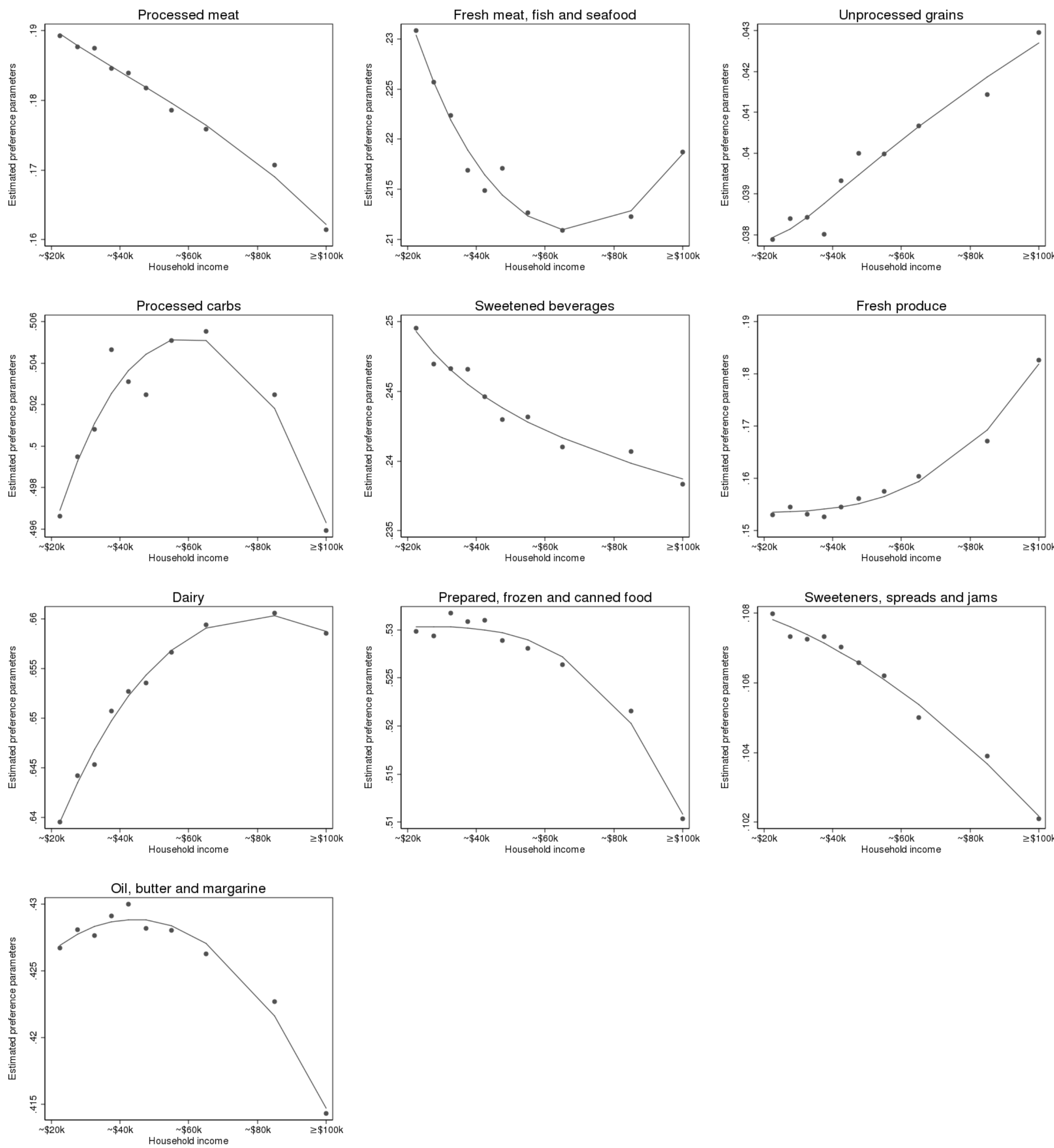




A.12 Preference Parameters and Demographics

I find that income is positively correlated with preferences for healthy products (this is consistent with the findings of Allcott et al. (2017) and Handbury (2013)). Two product categories deviate from this general pattern: processed carbohydrates and fresh meat and seafood. Specifically, preferences for processed meat, sweetened beverages, prepared, frozen and canned food, sweeteners spreads and jams to be decreasing with income. Preferences for oil, butter and margarine are slightly increasing with income for incomes under $40,000 and then decrease. Unprocessed grains and fruit and vegetables are increasing in income. Preferences for dairy increase with income except at the very top. Processed carbs and fresh meat and seafood are not monotonic with income – they are slightly concave and slightly convex, respectively.[82]

[82]Preferences for processed carbs increase with income for incomes under $60,000 and then decrease slightly. Symmetrically, preferences for fresh meat and seafood, decrease with income for incomes under $60,000 and then increase. However, both of these preference parameters have little variation with income.



Processed meat
Estimated preference parameters
.16 .17 .18 .19
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Fresh meat, fish and seafood
Estimated preference parameters
.21 .215 .22 .225 .23
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Unprocessed grains
Estimated preference parameters
.038 .039 .04 .041 .042 .043
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Processed carbs
Estimated preference parameters
.496 .498 .5 .502 .504 .506
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Sweetened beverages
Estimated preference parameters
.235 .24 .245 .25
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Fresh produce
Estimated preference parameters
.15 .16 .17 .18 .19
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Dairy
Estimated preference parameters
.64 .645 .65 .655 .66
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Prepared, frozen and canned food
Estimated preference parameters
.51 .515 .52 .525 .53
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Sweeteners, spreads and jams
Estimated preference parameters
.102 .104 .106 .108
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income
Oil, butter and margarine
Estimated preference parameters
.415 .42 .425 .43
~$20k ~$40k ~$60k ~$80k ≥$100k
Household income

A.12.1 Preferences and Education

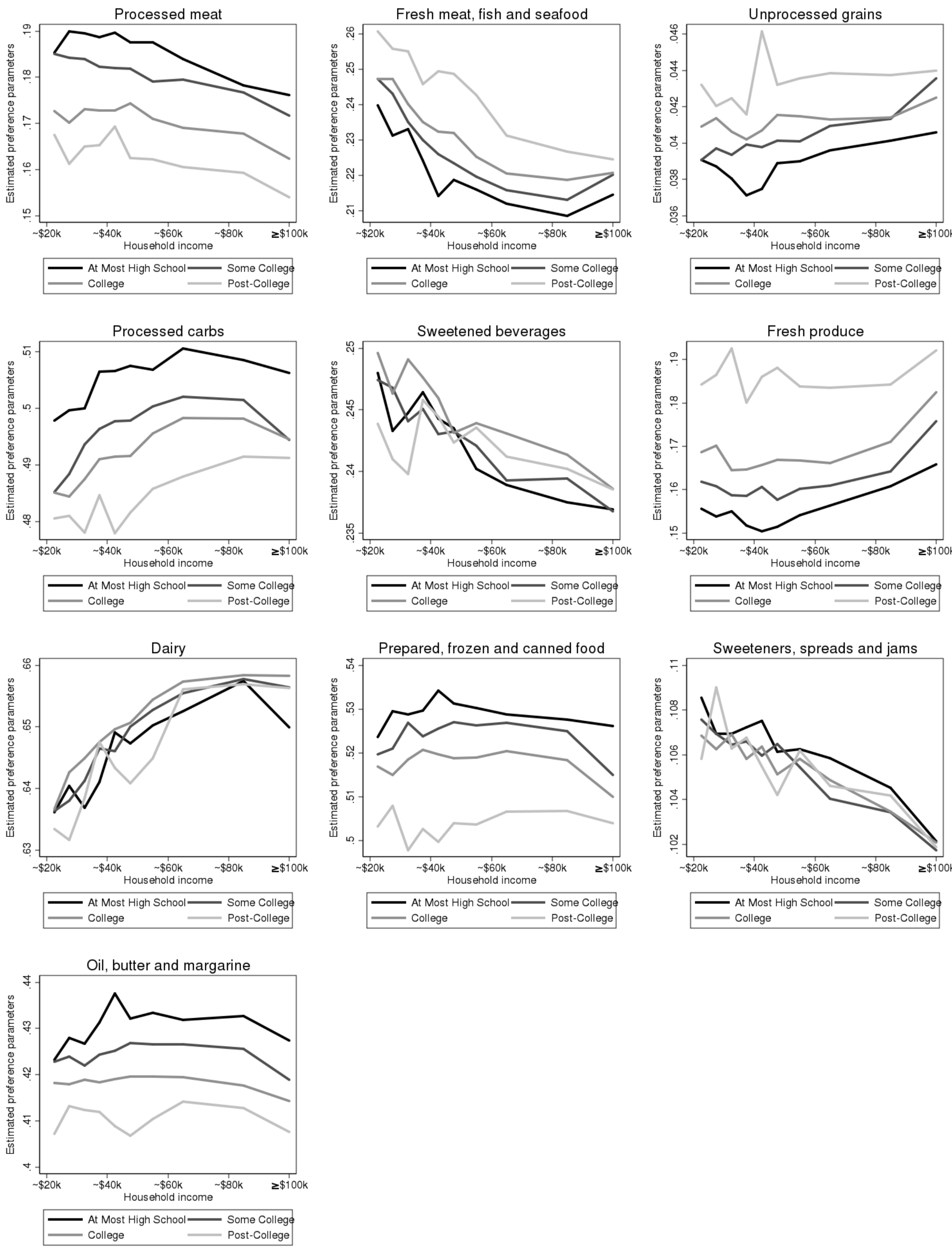


Table 14: *Regression coefficients: preference parameters on education and income*

(a) *Estimates for food products 1-5*

		1	2	3	4	5
education						
	$<$ college	-0.0041^{***} (0.0004)	0.0027^{***} (0.0006)	0.0014^{***} (0.0002)	-0.0078^{***} (0.0005)	-0.0002 (0.0004)
	college	-0.0140^{***} (0.0004)	0.0041^{***} (0.0006)	0.0024^{***} (0.0002)	-0.0087^{***} (0.0006)	0.0029^{***} (0.0004)
	$>$ college	-0.0213^{***} (0.0007)	0.0085^{***} (0.0011)	0.0052^{***} (0.0003)	-0.0124^{***} (0.0009)	-0.0036^{***} (0.0007)
income quartile						
	2	0.0009 (0.0005)	-0.0203^{***} (0.0008)	0.0001 (0.0002)	0.0088^{***} (0.0007)	-0.0027^{***} (0.0005)
	3	-0.0007 (0.0005)	-0.0262^{***} (0.0008)	0.0010^{***} (0.0002)	0.0103^{***} (0.0007)	-0.0069^{***} (0.0005)
	4	-0.0059^{***} (0.0006)	-0.0322^{***} (0.0009)	0.0025^{***} (0.0002)	0.0098^{***} (0.0008)	-0.0075^{***} (0.0006)
educ$\times$income						
$<$ college	2	-0.0020^{**} (0.0007)	0.0005 (0.0009)	0.0004 (0.0003)	0.0010 (0.0008)	0.0002 (0.0006)
	3	-0.0021^{**} (0.0007)	-0.0019^{*} (0.0010)	-0.0001 (0.0003)	0.0030^{***} (0.0009)	0.0025^{***} (0.0006)
	4	0.0005 (0.0007)	0.0003 (0.0010)	-0.0002 (0.0003)	0.0026^{***} (0.0009)	0.0016^{**} (0.0007)
college	2	-0.0008 (0.0007)	0.0049^{***} (0.0010)	0.0001 (0.0003)	-0.0035^{***} (0.0009)	-0.0013^{*} (0.0006)
	3	-0.0002 (0.0007)	0.0013 (0.0010)	-0.0001 (0.0003)	-0.0008 (0.0009)	0.0009 (0.0006)
	4	0.0039^{***} (0.0007)	0.0014 (0.0010)	-0.0015^{***} (0.0003)	0.0010 (0.0009)	-0.0003 (0.0007)
$>$ college	2	0.0019 (0.0010)	0.0109^{***} (0.0015)	-0.0002 (0.0004)	-0.0077^{***} (0.0013)	0.0029^{**} (0.0010)
	3	-0.0004 (0.0010)	0.0067^{***} (0.0014)	-0.0012^{**} (0.0004)	-0.0044^{***} (0.0012)	0.0066^{***} (0.0009)
	4	0.0035^{***} (0.0009)	0.0010 (0.0013)	-0.0023^{***} (0.0004)	0.0004 (0.0012)	0.0056^{***} (0.0009)

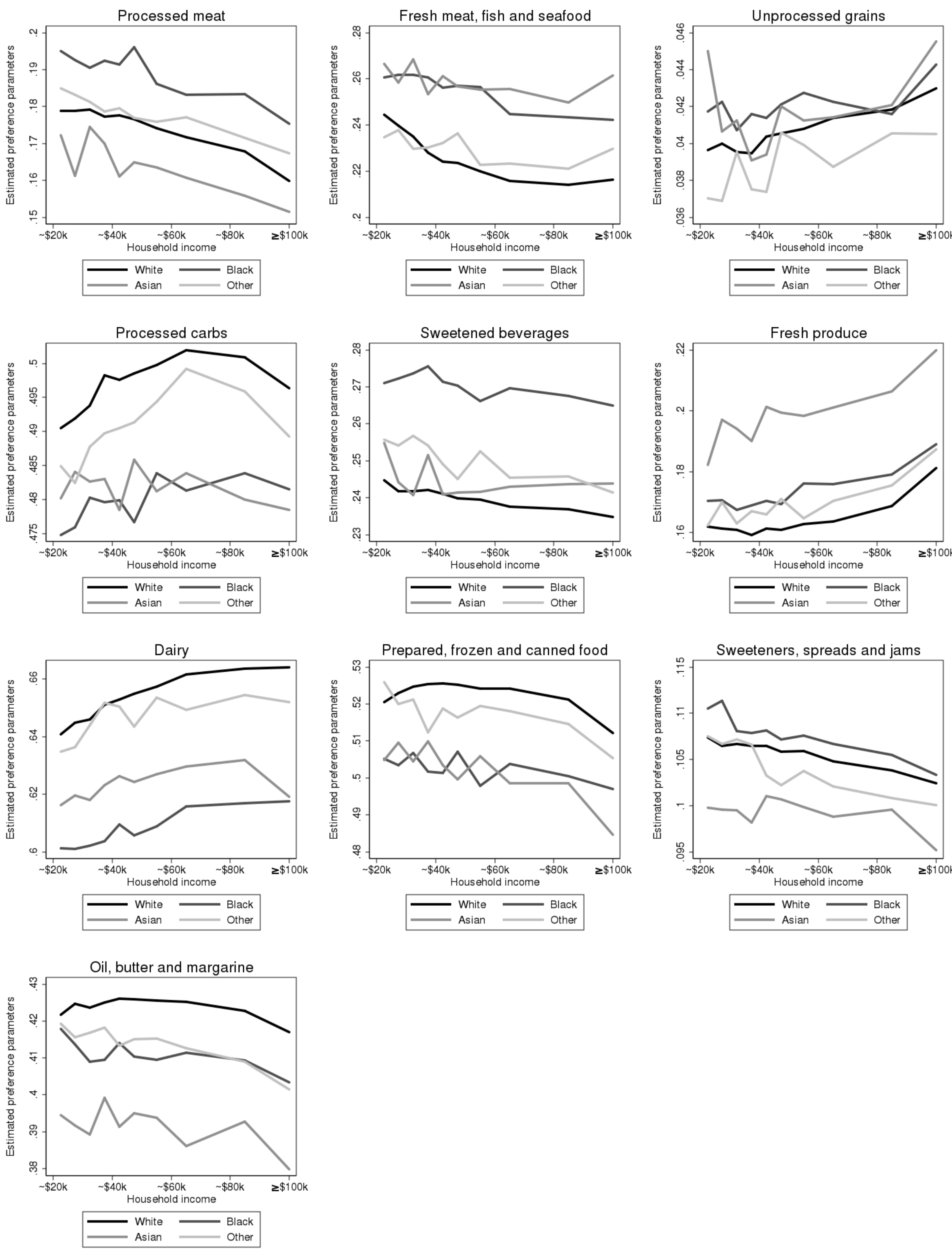
Standard error in parentheses
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
The regressions include zipcode-level income fixed effects.

(a) Estimates for food products 6-10

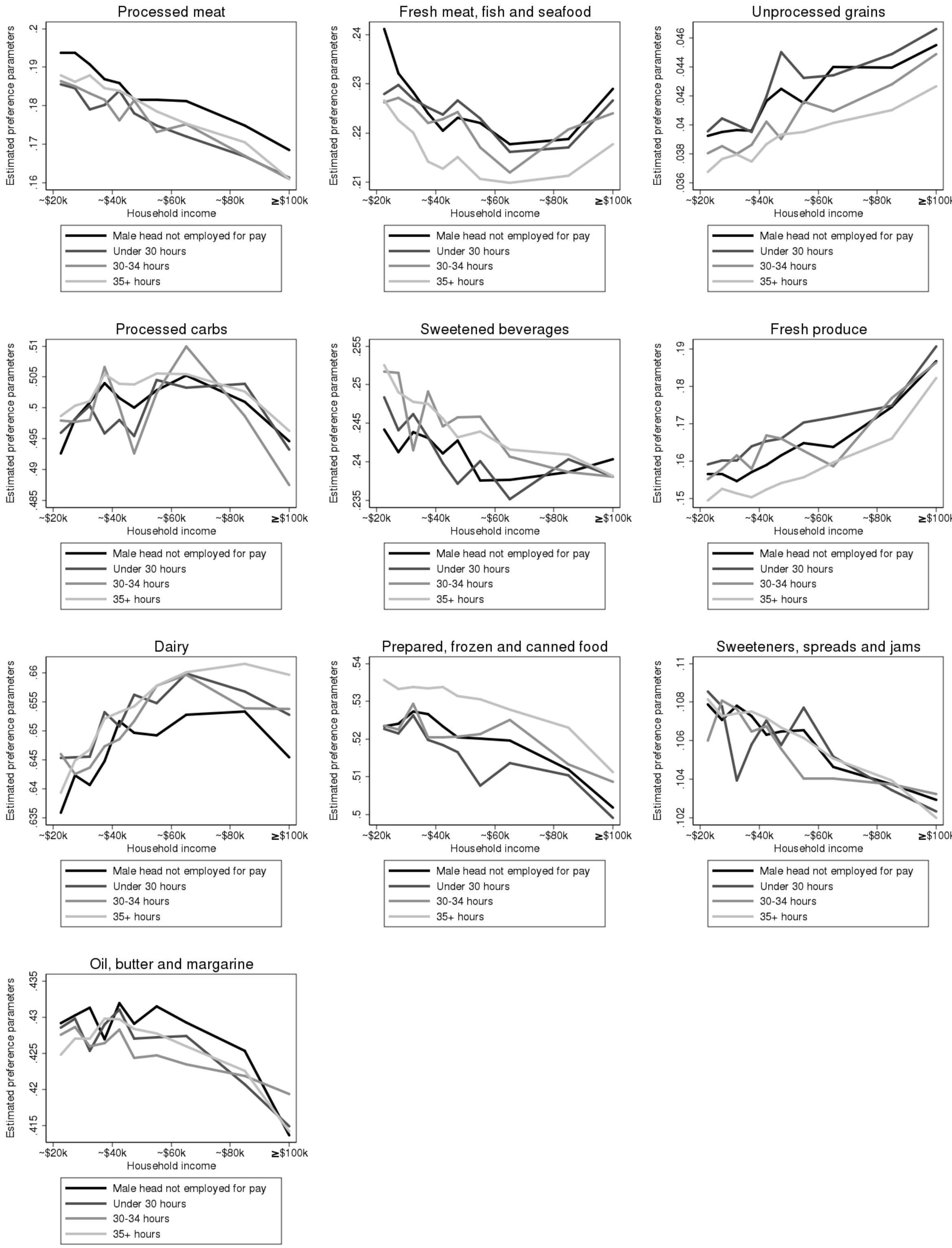
		6	7	8	9	10
education						
$<$ college		0.0034*** (0.0005)	0.0029*** (0.0005)	−0.0020*** (0.0005)	0.0001 (0.0002)	−0.0011** (0.0004)
college		0.0092*** (0.0005)	0.0057*** (0.0006)	−0.0077*** (0.0005)	−0.0007*** (0.0002)	−0.0059*** (0.0004)
$>$ college		0.0255*** (0.0008)	0.0051*** (0.0009)	−0.0191*** (0.0009)	−0.0005 (0.0003)	−0.0098*** (0.0007)
income quartile						
	2	−0.0042*** (0.0006)	0.0092*** (0.0007)	0.0084*** (0.0006)	−0.0015*** (0.0002)	0.0071*** (0.0005)
	3	−0.0014* (0.0006)	0.0155*** (0.0007)	0.0071*** (0.0007)	−0.0019*** (0.0002)	0.0067*** (0.0006)
	4	0.0024*** (0.0007)	0.0200*** (0.0008)	0.0059*** (0.0007)	−0.0032*** (0.0002)	0.0080*** (0.0006)
educ×income						
$<$ college	2	0.0016* (0.0008)	0.0010 (0.0008)	−0.0025*** (0.0008)	0.0001 (0.0003)	−0.0032*** (0.0007)
	3	−0.0001 (0.0008)	0.0009 (0.0009)	−0.0009 (0.0008)	−0.0007** (0.0003)	−0.0017* (0.0007)
	4	0.0008 (0.0008)	0.0003 (0.0009)	−0.0030*** (0.0008)	−0.0007* (0.0003)	−0.0049*** (0.0007)
college	2	0.0027*** (0.0008)	0.0007 (0.0009)	−0.0030*** (0.0008)	0.0006* (0.0003)	−0.0044*** (0.0007)
	3	0.0001 (0.0008)	0.0008 (0.0009)	−0.0020* (0.0008)	0.0009*** (0.0003)	−0.0030*** (0.0007)
	4	0.0004 (0.0008)	−0.0001 (0.0009)	−0.0026** (0.0008)	0.0007** (0.0003)	−0.0049*** (0.0007)
$>$ college	2	0.0003 (0.0012)	−0.0010 (0.0013)	−0.0053*** (0.0012)	0.0004 (0.0004)	−0.0066*** (0.0010)
	3	−0.0038*** (0.0011)	−0.0003 (0.0012)	−0.0019 (0.0012)	0.0010** (0.0004)	−0.0040*** (0.0010)
	4	−0.0062*** (0.0011)	−0.0005 (0.0012)	0.0015 (0.0011)	0.0010** (0.0004)	−0.0054*** (0.0009)

Standard error in parentheses
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
The regressions include zipcode-level income fixed effects.

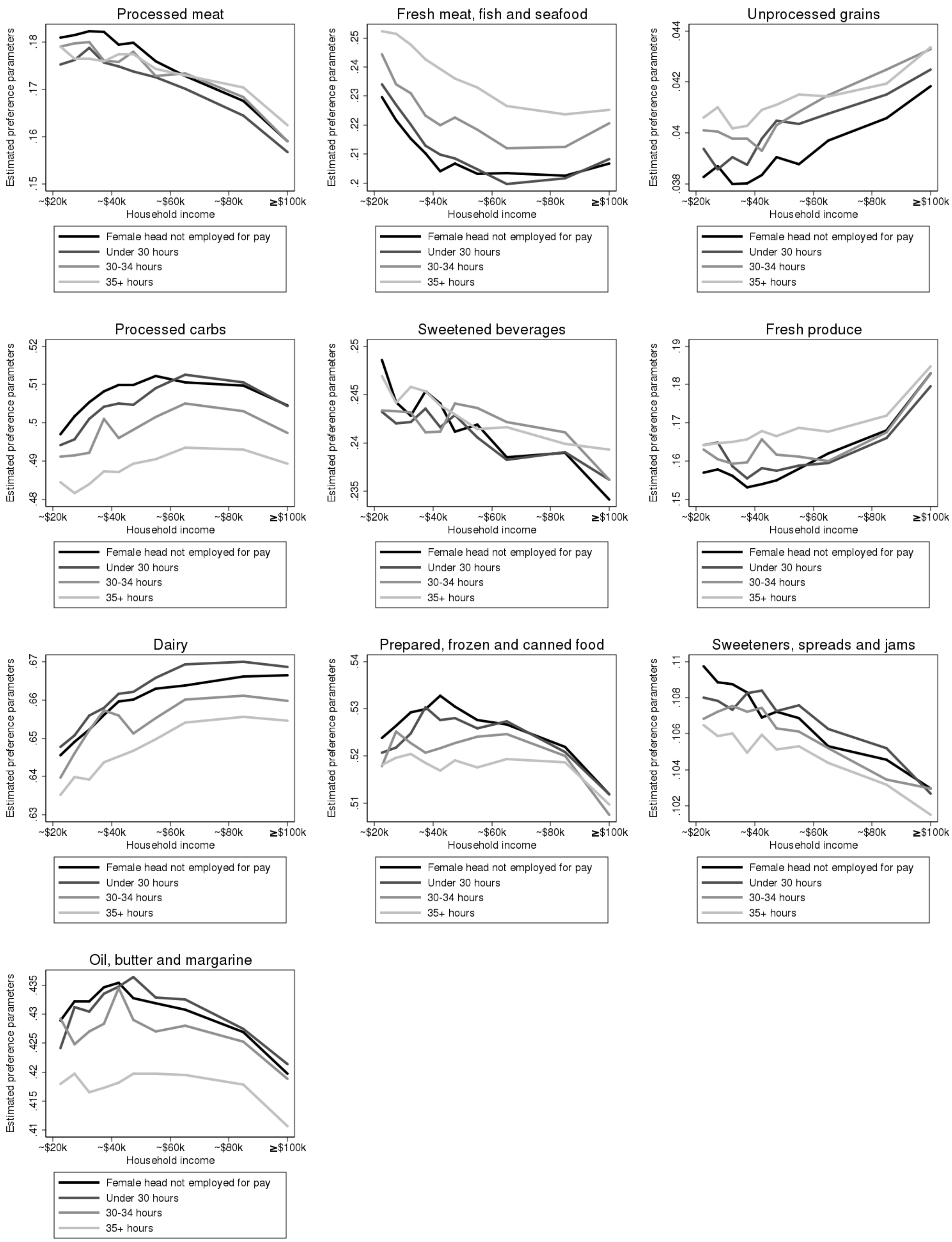
A.12.2 Preferences and Race



A.12.3 Preferences and Hours of Employment of the male head



A.12.4 Preferences and Hours of Employment of the female head



A.12.5 Preferences and Hours of Employment of both heads