THE EFFECT OF SUGAR-SWEETENED BEVERAGE CONSUMPTION ON CHILDHOOD OBESITY – CAUSAL EVIDENCE

Yan Yang

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Economics
Indiana University

October 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Joseph V. Terza, Ph.D., Chair

Charles Courtemanche, Ph.D.

May 18, 2016

Haeil Jung, Ph.D.

Henry Y. Mak, Ph.D.

Jisong Wu, Ph.D.

ACKNOWLEDGEMENTS

I would like to thank, first and foremost, my dissertation advisor, Dr. Joseph V. Terza, for his great help and guidance all the way along my doctoral journey. I have learnt many things from him, which would have significant impact for my future career. I am indebted to my research and dissertation committees, Dr. Charles Courtemanche, Dr. Haeil Jung, Dr. Henry Y. Mak, and Dr. Jisong Wu. Although it is hard to fully express my deep gratitude for their persistent support, I thank all of them for always opening their doors for me.

To Dr. Anne B. Royalty, who brought me to the Ph.D. program and has been consistently supporting me and encouraging me. To Dr. Mark Ottoni-Wilhelm, who helped me start to think as a heath economics researcher and taught me how to conduct a research study. I appreciate the knowledge and skills I gained from Dr. Steven Russell, Dr. Wendy Morrison, Dr. Paul S. Carlin, Dr. Subir K. Chakrabarti, Dr. Sumedha Gupta, Dr. Yaa Akosa Antwi, and Dr. Jaesoo Kim, and their detailed explanations to all my questions as well. I also would like to thank Dana M. Ward, our department secretory, who helped me get through many administrative hurdles during the process, especially when I was away from the campus.

To my husband Yue Fu and my dear parents, my motivation for pursuing this Ph.D. degree is from all of you. Thanks for your support during the past five years.

To my friends, my fellow classmates and many of those who I failed to mentioned in this short page. No matter where you were and where you are, I want you know that I cherish our friendship and will remember each single moment you shed light on my life. I appreciate every meeting and every interaction with each of you.

Yan Yang

EFFECT OF SUGAR-SWEETENED BEVERAGE CONSUMPTION ON CHILDHOOD OBESITY – CAUSAL EVIDENCE

Communities and States are increasingly targeting the consumption of sugar-sweetened beverages (SSBs), especially soda, in their efforts to curb childhood obesity. However, the empirical evidence based on which policy makers design the relevant policies is not causally interpretable. In the present study, we suggest a modeling framework that can be used for making causal estimation and inference in the context of childhood obesity. This modeling framework is built upon the two-stage residual inclusion (2SRI) instrumental variables method and have two levels – level one models children's lifestyle choices and level two models children's energy balance which is assumed to be dependent on their lifestyle behaviors.

We start with a simplified version of the model that includes only one policy, one lifestyle, one energy balance, and one observable control variable. We then extend this simple version to be a general one that accommodates multiple policy and lifestyle variables. The two versions of the model are 1) first estimated via the nonlinear least square (NLS) method (henceforth *NLS-based 2SRI*); and 2) then estimated via the maximum likelihood estimation (MLE) method (henceforth *MLE-based 2SRI*). Using simulated data, we show that 1) our proposed 2SRI method outperforms the conventional method that ignores the inherent nonlinearity [the linear instrumental variables (LIV) method] or the potential endogeneity [the nonlinear regression (NR) method] in obtaining the relevant estimators; and 2) the MLE-based 2SRI provides more efficient estimators (also consistent) compared to the NLS-based one. Real data analysis is conducted to illustrate the

implementation of 2SRI method in practice using both NLS and MLE methods. However, due to data limitation, we are not able to draw any inference regarding the impacts of lifestyle, specifically SSB consumption, on childhood obesity. We are in the process of getting better data and, after doing so, we will replicate and extend the analyses conducted here. These analyses, we believe, will produce causally interpretable evidence of the effects of SSB consumption and other lifestyle choices on childhood obesity. The empirical analyses presented in this dissertation should, therefore, be viewed as an illustration of our newly proposed framework for causal estimation and inference.

Joseph V. Terza, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Background and Significance

The US childhood obesity rate has risen from 5% in 1971-74 to 17% in 2009-10 (Anderson and Butcher, 2006; Ogden et al., 2012). This trend mirrors that of the adult obesity rate, which grew from 13% to 34% between 1960 and 2008 (Flegal et al., 1998; Ogden et al., 2010). The rise in obesity has become a leading public health concern, as adiposity contributes to health problems such as heart disease, diabetes, high blood pressure, and stroke (Sturm, 2002). This has prompted a growing number of policy proposals intended to reverse or slow the trend.

Sugar-sweetened beverages (SSBs), particularly soda, have become a popular target of such proposals, as soda is the single largest contributor to caloric intake (Block, 2004). Moreover, SSB calories may lead to larger increases in body weight than other sources of calories. A meta-analysis by Mattes (1996) finds that only 9% of calories from liquids are offset by subsequent downward adjustment in caloric intake, compared to 64% for solid foods. Additionally, SSBs have relatively high glycemic indices (Healthaliciousness.com, 2013).

Interventions to reduce SSB intake among children can take several forms. As of 2007, 34 U.S. states taxed soda sold in grocery stores while 39 states taxed soda sold in vending machines. However, the tax rates were all below 10% and the purpose was primarily to raise revenue (Levy et al., 2011; Fletcher et al., 2010b). In recent years, proposals for larger soda taxes at the federal, state, or local levels with the explicit purpose of curbing childhood obesity have become increasingly common (Fletcher et al., 2010b). A recent New York City law would have banned restaurants from selling sodas and other sugary beverages larger than 16 ounces, though the law was ultimately overturned by the

courts (Reuters, 2013). School-level policies, such as not having "pouring rights" to soda with bottling companies or prohibiting stores, snack bars, or vending machines from selling soda, are also increasingly common (Levy et al., 2011). New federal regulations on the nutritional content of foods and drinks sold in school vending machines are scheduled to take effect in the 2014-2015 academic year and should dramatically reduce SSB availability in schools (Shah, 2013).

Despite the growing popularity of SSB-related interventions, the case for singling out SSBs to reduce childhood obesity is largely based on evidence that is not causally interpretable. Vartanian et al. (2007) conduct a meta-analysis of 88 studies and find a positive association between soda intake and body weight. Malik et al. (2006) and Woodward-Lopez et al. (2011) reach similar conclusions after reviewing 25 and 56 observational studies, respectively. However, the associations produced by such observational studies may not reflect causal effects of soda on weight, in which case their relevance for policy is unclear. These associations could be driven partially or entirely by unobservable characteristics – such as an individual's level of interest in health – that might influence not only soda intake but also other determinants of weight (e.g. junk food consumption and exercise).[1] To the extent that the regressions do not control for these other determinants, the estimated effect of soda on weight could be exaggerated. Reverse causality is also a concern, as higher weight means greater caloric needs.

Perhaps because of the limited causal evidence upon which they are based, SSB-related interventions, as currently practiced, do not appear to have had clear effects on childhood obesity. Powell and Chaloupka (2009) only find evidence of an effect of state

[1] Such variables will be henceforth referred to as *confounders.*

soda taxes on adolescents' BMI among those at risk of overweight, while Sturm et al. (2010) estimate a negative but modest relationship between soda taxes and BMI among fifth graders. Fletcher et al. (2010b) study a longer time period and a broader age range (3-18) than these prior studies, and more thoroughly account for omitted variables by including state fixed effects. Changes in state soda tax rates are not significantly associated with changes in child BMI, overweight, or obesity, as the decrease in calories from soda is offset by an increase in calories from whole milk. Finally, Forshee et al. (2005), Fletcher et al. (2010a), and Taber et al. (2011) find no evidence of an effect of removing SSBs or junk foods from school vending machines on child BMI. While such retrospective program evaluations are useful, trial and error can be an expensive way to gain information about which SSB-related interventions best combat obesity. Some possible interventions, such as educational programs, impose large fiscal costs. Others, including taxes and restrictions, are not fiscally costly but economic theory suggests they would result in net social costs unless they reduce weight.

An alternative approach is to gather prospective evidence through small-scale pilot experiments. [2] James et al. (2004) show that a randomized nutrition education program among 29 elementary school classes in the United Kingdom reduced carbonated drink consumption and overweight and obesity rates. Ebbeling et al. (2006) show that a home-based randomized experiment, which featured counseling and weekly deliveries of non-caloric beverages among 103 13-18 year olds in Boston, only significantly reduced BMI among the heaviest teenagers. Sichieri et al. (2009) randomized 47 4th grade classes in

[2] The discussion is limited to randomized experiments among children that included weight-related outcomes. Other experiments attempt to randomize soda intake among adults or focus on only intermediate outcomes such as dietary habits. See Levy et al. (2011) and Woodward-Lopez et al. (2011) for discussions of these studies.

Brazil to an educational program focused on reducing carbonated sugar-sweetened beverages, finding that the intervention led to a substitution from soda to juice, with body mass index (BMI) only dropping among overweight children. While these randomized experiments provide some causally-interpretable evidence that SSB-related interventions can reduce childhood obesity, their generalizability is limited by their small samples and the fact that only one occurred in the U.S. Accumulating a large enough evidence base from randomized experiments to motivate large-scale U.S. policy would be expensive and time consuming.

Therefore, we propose an approach designed to produce: 1) causally-interpretable evidence on the impact of SSB consumption on children's weight; and 2) quantitative recommendations for potential SSB-related childhood obesity-fighting policies aimed at specified energy balance goals. Our approach requires only observational data, avoiding the large costs of trial interventions and randomized experiments. Unlike conventional methods, it explicitly accounts for inherent nonlinearity and the potential endogeneity of relevant behaviors in the modeling of child energy balance.

Our econometric framework comprises two components. First, we model children's lifestyle choices that contribute to energy balance as a series of nonlinear regression equations, referred to as *lifestyle regressions*, with dependent variables such as calories from SSBs, calories from other sources, and minutes of physical activity per day. The main independent variables in this lifestyle regression system are the observable analogs of "prospective policy interventions", such as the prices of SSBs, other foods and drinks, and fast-food meals; access to fast-food restaurants, full-service restaurants, grocery stores, Walmart Supercenters, and warehouse clubs; and nutrition information spending. These

variables all relate to potential policy levers: the price variables to taxes and subsidies; the establishment variables to taxes, subsidies, and moratoria for particular types of businesses; and nutritional education funding to further information spreading efforts. In the second component, we specify a regression equation, referred to as the *energy balance regression*, whose dependent variable is a measure of children's energy balance (BMI percentile (%-ile), body fat percentage (%), and definitional overweight and obesity), and whose key independent variables are the lifestyle variables, indicating children's eating and exercise habits. The obesity-related lifestyle choices may be endogenous due to some unobservable variables relating to health status, genetics, parental characteristics, etc. that impact both a child's energy balance and his or her eating or exercise habits. We correct for this potential endogeneity bias by implementing the two-stage residual inclusion (2SRI) instrumental variables method suggested by Terza et al. (2008), where instruments are the policy related variables mentioned above in the context of the first component of the econometric framework. The 2SRI method is particularly appropriate in this context because it is designed to account for the inherent nonlinearity of both components of the model. Based on this modeling framework, we are able to improve the evidence base on possible SSB-related interventions for combatting childhood obesity: combining the results from both components of the regression system, we estimate the causally-interpretable effects of the SSB-related prospective policy levers on energy balance. We also provide a way to estimate the change in a prospective policy lever that would be required to achieve a desired energy balance outcome, which is different from a commonly used linear approximation approach.

The remainder of the dissertation is organized as follows. Chapter 2 discusses a simplified version of the model that includes only one policy, one lifestyle, one energy balance, and one observable control variable, and derives the estimators that can be used to evaluate policy effects and provide policy recommendations. As a comparison, estimators based on the conventional methods that ignore nonlinearity or endogeneity are also provided. In Chapter 3, the simple model in Chapter 2 is extended to be a general model that accommodates multiple policy and lifestyle variables and, correspondingly, more general policy effect estimators are derived. Chapter 4 introduces a full information version of the simple model by assuming known forms for the conditional probability density functions of the lifestyle variable (soda calorie intake) and for the energy balance variable (body fat %), and incorporating this information into the estimation of the relevant parameters. By doing this, we expect to obtain more efficient estimators. Using the same logic, chapter 5 discusses the full information version of the general model. The performance of the model introduced in chapter 2-5 are examined using simulated data (i.e. chapter 2 and 4) or real data (i.e. chapter 3 and 5). Finally, Chapter 6 summarizes and discusses the models and results put forth in Chapters 2 through 5.

Chapter 2: A Simplified Version of the Model and Proposed Empirical Policy Analytic Methods

The estimation of the causal effects of childhood behaviors (e.g. SSB calories intake, other calories intake, exercise, etc.) on energy balance is complicated by the fact that there might be some unobserved characteristics (e.g genetics or quality time with parents) that correlate with both weight and these behaviors. Failure to control for such unobserved confounding factors relegates conventional regression-based estimates to interpretation as merely indicative of statistical association, supplying little or no useful content for policy makers. The modeling framework proposed here takes explicit control of these factors so that the statistical estimates that it produces will be causally interpretable and, therefore, relevant to policy analysts and policy makers. This model has two levels: the first level models the effects of exogenous changes in the potential policy variables on children's weight-related behaviors; the second level focuses on the causal effects of changes in these behaviors on child energy balance. Because child energy balance regressions in the second level are inherently nonlinear [proportional regressions for BMI %-ile and body fat % (see Basu and Manca, 2012; Buis, et al., 2012; and Paolino, 2001); logit analyses for obesity and overweight (see Wooldridge, 2010, Chapter 15)], we propose the use of the two-stage residual inclusion (2SRI) instrumental variables method suggested by Terza et al. (2008) to address the endogeneity bias due to unobserved confounders in nonlinear models.

To keep things simple without loss of generality, we start with the discussion of a simplified version of the model that includes only one energy balance variable (body fat %), one policy variable (soda price), one lifestyle variable (soda calorie consumption), and one

observable control variable (age). In this simple illustrative example, we demonstrate the corresponding two-level econometric framework and the way to consistently estimate the relevant coefficients. We also derive estimators that can be used to evaluate policy (soda price or other policies aimed at affecting soda consumption directly) effects and provide recommendations for policy changes (soda price change) or lifestyle changes (soda calorie consumption change) aimed at achieving a desired energy balance outcome (ideal population mean of body fat %). As a comparison, estimators based on conventional methods that ignore nonlinearity [the simple linear instrumental variables (LIV) method] or endogeneity [the nonlinear regression (NR) method] are also provided, and such comparison is made using the simulated data.

2.1 Regression Representation of the Simplified Model

We posit the following lifestyle regression model

$$L = \exp(X_o \alpha_o + P\alpha_p) + X_u \tag{2.1}$$

where

L ≡ daily soda calories consumed (cal.)

$X_o \equiv [1 \quad AGE]$

AGE ≡ children's age (years)

P ≡ soda price ($ per 2 liters)

$E[L|\, X_o, P] = \exp(X_o \alpha_o + P\alpha_p)$; X_u is the regression error term; and $\alpha' \equiv [\alpha_o' \quad \alpha_p]$ is the vector of parameters to be estimated with $\alpha_o' \equiv [\alpha_{CONST} \quad \alpha_{AGE}]$. In addition, we assume that

$$E[EB \mid L, X_o, X_u] = \Lambda(L\beta_L + X_o\beta_o + X_u\beta_u) \tag{2.2}$$

where $\Lambda(\quad)$ is the logistic cumulative distribution function (cdf) and $\beta' = [\beta_L \quad \beta_o' \quad \beta_u]$ is the vector of parameters to be estimated with $\beta_o' \equiv [\beta_{CONST} \quad \beta_{AGE}]$. This yields the following form for the energy balance regression model

$$EB = \Lambda(L\beta_L + X_o\beta_o + X_u\beta_u) + e \tag{2.3}$$

where $e = EB - \Lambda(L\beta_L + X_o\beta_o + X_u\beta_u)$ is the regression error term. The regression model in (2.1) and (2.3) accounts for the potential endogeneity of soda calorie consumption (L) through the explicit inclusion of its unobserved confounders, X_u, in the energy balance equation. The observable control variable (AGE) included in X_o is assumed to be exogenous in both equations. We apply the two-stage residual inclusion (2SRI) method suggested by Terza et al. (2008) to obtain estimates of the αs and the βs in the model. The 2SRI method requires at least one instrumental variable that is highly correlated with soda calorie consumption, L, but correlated with body fat %, EB, only through its influence on soda calorie consumption. Soda price (P) satisfies this condition, and can be used as the instrumental variable. The two stages of the 2SRI method are:

Stage 1 – use the nonlinear least squares (NLS) method to estimate the lifestyle regression (2.1) and obtain consistent estimators, $\tilde{\alpha}_o^{2SRI}$ and $\tilde{\alpha}_P^{2SRI}$, then calculate $\tilde{X}_u^{2SRI} = L - \exp(X_o\tilde{\alpha}_o^{2SRI} + P\tilde{\alpha}_P^{2SRI})$;

Stage 2 – obtain the consistent estimators, $\tilde{\beta}_L^{2SRI}$, $\tilde{\beta}_o^{2SRI}$, and $\tilde{\beta}_u^{2SRI}$ by applying the NLS method to the following version of (2.3)

$$EB = \Lambda(L\beta_L + X_o\beta_o + \tilde{X}_u^{2SRI}\beta_u) + e^{2SRI} \tag{2.4}$$

with $\tilde{X}_u^{2SRI}$ obtained from the stage 1.

The 2SRI method described above takes account of both the inherent nonlinearity and the potential endogeneity of soda calorie consumption in the modeling of child energy balance and, as a result, all the estimates of the αs and the βs are consistent.

2.2 Average Incremental Effects in the Simplified Model

Using the 2SRI parameter estimates and the corresponding lifestyle and energy balance equations, we can estimate the effect of soda calorie consumption or soda price on body fat %. We first derive estimators of the change in body fat %, on average, in response to a particular change in soda calorie consumption or soda price, based on the 2SRI method. Following the approach of Terza and Wu (2016), using the 2SRI parameter estimates, the average incremental effect (AIE) of an exogenous policy-driven increment in soda calorie consumption, say Δ_L, on body fat % can be estimated as

$$\left(\sum_{i=1}^{n}\frac{1}{n}\widetilde{EB_i(\Delta_L)}\right) - \overline{EB} \tag{2.5}$$

where

$$\widetilde{EB_i(\Delta_L)} = \Lambda((L_i + \Delta_L)\tilde{\beta}_L^{2SRI} + X_{oi}\tilde{\beta}_o^{2SRI} + \tilde{X}_{ui}^{2SRI}\tilde{\beta}_u^{2SRI})$$

$\tilde{X}_{ui}^{2SRI} = L_i - \exp(X_{oi}\tilde{\alpha}_o^{2SRI} + P_i\tilde{\alpha}_P^{2SRI})$; the i subscript refers to the ith sample member; $\tilde{\alpha}^{2SRI}$s and $\tilde{\beta}^{2SRI}$s are the 2SRI estimators; and $\overline{EB}$ denotes the sample average for body fat %. Similarly, the estimated AIE of an exogenous increment in soda price by the amount Δ_P on body fat % is[3]

$$\left(\sum_{i=1}^{n}\frac{1}{n}\widetilde{EB_i(\Delta_P)}\right) - \overline{EB} \qquad (2.6)$$

where

$$\widetilde{EB_i(\Delta_P)} = \Lambda(\tilde{\tilde{L}}_i(\Delta_P)\tilde{\beta}_L^{2SRI} + X_{oi}\tilde{\beta}_o^{2SRI} + \tilde{X}_{ui}^{2SRI}\tilde{\beta}_u^{2SRI})$$

$$\tilde{\tilde{L}}_i(\Delta_P) = \exp(X_{oi}\tilde{\alpha}_o^{2SRI} + (P_i + \Delta_P)\tilde{\alpha}_P^{2SRI}) + \tilde{X}_{ui}^{2SRI}$$

and $\tilde{X}_{ui}^{2SRI}$ and $\overline{EB}$ are defined as in equation (2.11). It is easy to show that the estimated AIEs in equations (2.5) and (2.6) are consistent as the coefficient estimates ($\tilde{\alpha}^{2SRI}$s and $\tilde{\beta}^{2SRI}$s) used are consistent. We will refer to these two AIE estimators as the 2SRI-estimated AIEs.

[3] For detailed derivations of equation (2.5) – (2.6) see Appendix 2C.

2.3 Requisite Policy Changes to Achieve an EB Target in the Simplified Model

In addition to the aforementioned AIE analyses, we also provide other empirical measures that may be of interest to policy makers. For example, we derive an estimate of the requisite change in soda calorie consumption (soda price) for achieving a pre-specified energy balance target. Such a measure would be essential to the design of a policy intervention. Let EB^0 be the average ideal level of body fat %. Then the estimated change in soda calorie intake that would be required to bring the current average level of body fat %, $\overline{EB}$, down to the ideal one, EB^0, can be obtained by solving

$$EB^0 = \sum_{i=1}^{n} \frac{1}{n} \widetilde{EB_i(\tilde{\Delta}_L^0)} \tag{2.7}$$

where

$$\widetilde{EB_i(\tilde{\Delta}_L^0)} = \Lambda((L_i + \tilde{\Delta}_L^0)\tilde{\beta}_L^{2SRI} + X_{oi}\tilde{\beta}_o^{2SRI} + \tilde{X}_{ui}^{2SRI}\tilde{\beta}_u^{2SRI}),$$

$\tilde{X}_{ui}^{2SRI} = L_i - \exp(X_{oi}\tilde{\alpha}_o^{2SRI} + P_i\tilde{\alpha}_P^{2SRI})$; the $\tilde{\alpha}$s and the $\tilde{\beta}$s are the coefficient estimates obtained from 2SRI; and $\tilde{\Delta}_L^0$ denotes the estimated change in L required to achieve EB^0 and it is the only unknown value in this equation. Similarly, the estimated change in soda price that would be necessary to bring $\overline{EB}$ down to EB^0 is $\tilde{\Delta}_P^0$, such that

$$EB^0 = \sum_{i=1}^{n} \frac{1}{n} \widetilde{EB_i(\tilde{\Delta}_P^0)} \tag{2.8}$$

where

$$\widetilde{EB_i(\tilde{\Delta}_P^0)} = \Lambda(\tilde{\tilde{L}}_i(\tilde{\Delta}_P^0)\tilde{\beta}_L^{2SRI} + X_{oi}\tilde{\beta}_o^{2SRI} + \tilde{X}_{ui}^{2SRI}\tilde{\beta}_u^{2SRI})$$

$$\tilde{\tilde{L}}_i(\Delta_P) = \exp(X_{oi}\tilde{\alpha}_o^{2SRI} + (P_i + \Delta_P)\tilde{\alpha}_P^{2SRI}) + \tilde{X}_{ui}^{2SRI}$$

and can be obtained by solving equation (2.8). There is no closed form solution for $\tilde{\Delta}_L^0$ (or $\tilde{\Delta}_P^0$) as equation (2.7) [or (2.8)] is nonlinear. In practice, we use Stata/Mata Optimize procedure to approximate it by determining the value that minimizes the squared difference between EB^0 and $\sum_{i=1}^{n} \frac{1}{n} \widetilde{EB_i(\tilde{\Delta}_L^0)}$ [or $\sum_{i=1}^{n} \frac{1}{n} \widetilde{EB_i(\tilde{\Delta}_P^0)}$].[4] Based on equation (2.7) [or (2.8)], we can obtain the 2SRI-based $\tilde{\Delta}_L^0$ (or $\tilde{\Delta}_P^0$) by replacing the $\tilde{\alpha}$s and $\tilde{\beta}$s with $\tilde{\alpha}^{2SRI}$s and $\tilde{\beta}^{SRI}$s.

We prefer the approach discussed above in obtaining the relevant policy recommendation estimators, although it is a little complicated in calculation. Typically, for simplicity, researchers use a more conventional approach that relies on linear approximation. They simply divide the difference between the targeted average body fat % (EB^0) and the current average body fat % ($\overline{EB}$) by the estimated AIE with the relevant causal increment set equal to 1. Theoretically, there is no difference between these two approaches in obtaining estimated required changes in L or P if the regressions used in the method have linear functional forms. But when the regressions are nonlinear, the linear approximation approach may yield unreliable estimates that differ substantially from the true values. In the model we propose, the relevant relationships are inherently nonlinear. So the conventional approach based on linear approximation is not appropriate. This can be illustrated by Figure 2.1.A and 2.1.B. In Figure 2.1.A, we draw the true response curve of the AIE on body fat % for different values of Δ_L.[5] The response curve based on the

[4] The corresponding Stata program is available upon request.
[5] We plot the value of the true AIE(Δ_L) over varying Δ_L (ranges from -100 to 100, with 0.5 as the increment) in STATA, and find the true response curve is convex and passes through the origin.

conventional approach (referred to as the "linear approximation to the true response curve" in the figure) is linear and intersects with the true response curve when Δ_L is 1.[6] Both response curves are increasing as more soda calorie consumption indicates higher body fat %. The targeted level of average body fat % (15%) is lower than the current level (23%), so the expected change in body fat % is negative and we can draw a horizontal line [denoted by $\text{AIE}(\Delta_L^0)$ in the figure] below the x-axis to find out the change in L necessary to achieve the expected decrease in body fat %. The intersection of this horizontal line with the true response curve will give us the true requisite change in L (Δ_L^0), while the intersection of this horizontal line with the linear approximation curve will give us the change in L obtained from the conventional approach (Δ_L^{CONV}). Clearly, Δ_L^0 is greater than Δ_L^{CONV} in magnitude, indicating that the conventional approach tends to underestimate the actual decrease in soda consumption necessary to bring the average body fat % down to the ideal level. Similarly, we draw another figure to show the relationship between the change in P and the corresponding AIE on body fat % (see Figure 2.1.B). The true response curve is decreasing as higher soda price indicates lower body fat %.[7] The response curve based on the conventional approach (referred to as the "linear approximation to the true response curve" in the figure) is linear and decreasing, and intersects with the true response curve when Δ_P is 1.[8] After drawing a horizontal line [denoted by $\text{AIE}(\Delta_P^0)$ in the figure] below

[6] To obtain this linear approximation curve, we first find out the point in the graph that indicates the true AIE of one unit increase in L on body fat %, and this point should coincide with the one in the true response curve; we then draw a straight line, which is the line we are looking for, that passes through this point and the origin. Therefore, the two response curves in Figure 2.1.A should intersect at $\Delta_L = 1$.

[7] We plot the value of the true AIE(Δ_P) over varying Δ_P (ranges from 0 to 2, with 0.01 as the increment) in STATA, and find the true response curve is convex and passes through the origin.

[8] To obtain this linear approximation curve, we first find out the point in the graph that indicates the true AIE of one unit increase in P on body fat %, and this point should coincide with the one in the true response curve;

the x-axis [and above the line AIE(1)],[9] we obtain the true requisite change in P (Δ_P^0) and the change in P based on the conventional approach (Δ_P^{CONV}). Then we can see that Δ_P^0 is smaller than Δ_P^{CONV}, which means that the conventional approach tends to overestimate the actual increase in soda price necessary to bring the average body fat % down to the ideal level in this case.[10]

2.4 Simulation Comparison of Estimators in the Context of the Simplified Model

In the above three sections, i.e. section 2.1-2.3, we discussed our modeling framework, in the context of the simple case, and the way to obtain consistent coefficient estimates, AIE and policy recommendation estimators. To better illustrate our idea and examine our proposed method, we conduct a simulation study. Specifically, we compare our method to the conventional methods that ignore inherent nonlinearity [the simple linear instrumental variables (LIV) method] or potential endogeneity [the nonlinear regression (NR) method] when modeling energy balance outcomes, i.e. body fat % in this case. In this section, we first describe our sampling design, and then discuss the coefficient, AIEs, and policy recommendation estimators obtained from each of the three methods, i.e. 2SRI, LIV, and NR, using the simulated data.

---

we then draw a straight line, which is the line we are looking for, that passes through this point and the origin. Therefore, the two response curves in Figure 2.1.B should intersect at $\Delta_P = 1$.

[9] We put this horizontal line above the line AIE(1) in Figure 2.1.B because we would like to explain the results in Table 2.5.B, in which the true Δ_P^0 is less than 1. This horizontal line can also be put below the line AIE(1) if the targeted decrease in the average body fat % is so large such that the increase in P is expected to be greater than 1. In this case, $\Delta_P^0 > \Delta_P^{CONV}$

[10] Δ_P^0 may be greater than Δ_P^{CONV}, also see footnote 9.

2.4.1 Sampling Design in the Simplified Model

We simulate a sample of 200,000 children (ages 2-19) with four variables: children's age (years), daily soda calories consumed (cal.), body fat %, and soda price ($ per 2 liters).[11] We first generate data for age and soda price based on pre-specified values of the means and variances,[12] and then, generate data for daily soda calorie consumption. Because soda calorie consumption is nonnegative, we assume it to be a Generalized Gamma (GG) [three parameter] variate which has the following probability density function (pdf) [Manning et al., 2005]

$$f(L \mid X_o, P; \kappa, \mu, \sigma) = \frac{\gamma^{\gamma}}{\sigma L \sqrt{\gamma}\Gamma(\gamma)} \exp(Z\sqrt{\gamma} - V) \qquad L \geq 0 \qquad (2.9)$$

where L is the lifestyle variable, soda calorie consumption; κ, μ and σ are the basic parameters of the distribution; Γ( ) is the gamma function; $\gamma = |\kappa|^{-2}$; $Z = \text{sign}(\kappa)\{\ln(L) - \mu\}/\sigma$; $V = \gamma \exp(|\kappa| Z)$; $\mu = X_o \alpha_o + P\alpha_p$, X_o is a vector that consists of the observable control variable, age, and a constant term, P is the policy variable, soda price, and the αs are the parameters .

We make this assumption because the GG distribution subsumes many different distributions that are commonly used for non-negative random variables, such as Weibull, Exponential, Log-normal, and so on. By doing this, we are safe to argue that our simulation results are not limited to a specified non-negative distribution, i.e. Weibull, Exponential, Log-normal, etc., for L.

[11] We use Stata/Mata to generate the simulated sample based on equation (2.9) - (2.14) discussed below. The corresponding Stata program is available upon request.
[12] Age and soda price are assumed to be uniformly distributed with pre-specified means and variances.

To generate L, we need to know its cumulative distribution function (cdf). According to Stacy and Mihram (1965), the conditional cdf of the GG variable, L, is

$$F(L \mid X_o, P; a, v, p) = \begin{cases} \dfrac{\gamma(v, (L/a)^p)}{\Gamma(v)} & \text{if} \quad p > 0 \\ 1 - \dfrac{\gamma(v, (L/a)^p)}{\Gamma(v)} & \text{if} \quad p < 0 \end{cases} \tag{2.10}$$

where a, v, p are the basic parameters in this specification; and $\gamma(b, c)$ denotes the incomplete gamma function defined as $\gamma(b, c) = \int_0^c t^{b-1} e^{-t}\, dt$. The parameterization in this case differs from the one under which we specify the pdf in (2.9). Manning et al. (2005) provides a crosswalk between the form in (2.9) and the Stacy and Mihram parameterization. Therefore, we can express a, v, and p as functions of κ, μ, and σ, i.e. $a = \dfrac{\exp(\mu)}{\left(\dfrac{1}{|\kappa|^2}\right)^{\frac{\sigma}{\kappa}}}$, $v = \dfrac{1}{|\kappa|^2}$, and $p = \dfrac{\kappa}{\sigma}$. When $p > 0$, we have

$$L = a\ \gamma^{-1}(v, \Gamma(v)\, U[0,1])^{\frac{1}{p}} \tag{2.11}$$

where $\gamma^{-1}(d, j)$ denotes the inverse incomplete gamma function defined such that if $j = \gamma(d, k)$ then $k = \gamma^{-1}(d, j)$; U[0, 1] denotes the uniform random variable on the unit interval. Based on (2.11), we can now generate data for L by picking values for κ, σ, and

the αs, and hence calculating a, v, and p, which are the parameters used in (2.11).[13] After obtaining L, X_u can be generated as

$$X_u = L - \exp(X_o \alpha_o + P\alpha_p + C) \tag{2.12}$$

where $C = \ln[\kappa^{2\sigma/\kappa} C^*]$, $C^* = \Gamma\{(1/\kappa^2) + (\sigma/k)\}/\Gamma\{1/\kappa^2\}$;[14] and all the unobserved factors, other than age and soda price, that affect soda calorie consumption are captured by X_u. With the generated value of X_u in hand we are able to generate data for body fat %. Because the value of body fat % ranges from 0 to 1,[15] we assume it to be beta distributed (Basu and Manca, 2012; Buis, et al., 2012; Paolino, 2001), and the corresponding pdf is

$$h(EB \mid L, X_o, X_u; \xi, \mu) = \frac{\Gamma(\xi)}{\Gamma(\xi\mu)\Gamma(\xi(1-\mu))} EB^{\xi\mu-1}(1 - EB)^{\xi(1-\mu)-1} \qquad 0 < EB < 1 \tag{2.13}$$

where EB represents the energy balance variable (body fat %); $\mu = E[EB \mid L, X_o, X_u] = \Lambda(L\beta_L + X_o\beta_o + X_u\beta_u)$, $\Lambda(\ \ )$ is the logistic cdf; and ξ and the βs are the parameters. To generate EB, we pick values for ξ and the βs, and use the inverse transform method (Ross, 1997, p. 62)

$$EB = H^{-1}(U[0,1]) \tag{2.14}$$

[13] We cannot get the inverse incomplete gamma function directly in Stata/Mata, but we can get it indirectly. For the details of the data generation of L, see Appendix 2A.
[14] See Manning et al. (2005) for the conditional mean of the Generalized Gamma random variable.
[15] In the simulated data, body fat % is measured as a decimal instead of a percentage.

where $H^{-1}(\ )$ is the inverse beta cdf, and U[0,1] denotes the uniform random variable on the unit interval.[16]

Note that in this design, EB depends on the unobserved component, X_u, which is clearly correlated with L [see equation (2.12)]. Therefore, L is endogenous in (2.13) if β_u is nonzero. In fact, in constructing simulated samples, we can control the degree of endogeneity by varying the absolute value of β_u in the sampling design.

To make the simulation results more informative, we chose the relevant parameters for the sampling design so as to be as realistic as possible for children aged 2-19 based on the literature and some online resources. We set the mean soft drink price per 2 liters at $1.35, and mean age at 10.5 years. We also adjusted the relevant parameters to make the mean of daily soda calories consumed around 120 cal. and the mean body fat % around 23%.[17] Attention was also given to other important aspects of the relevant distributions. For example, body fat % is seldom close to 0 or 1, so the tails of the distribution in the simulated sample should be relatively thin. Similarly, soda calories consumed per day may not exceed 3200 cal. as this is the approximate maximum amount of calories needed daily for an active male aged 14-18 years.[18] Moreover, the signs of the parameters (the αs and the βs) were chosen so as to be meaningful (e.g. α_P is negative, in keeping with the law of demand; and β_L is positive, as increased soda consumption is likely to lead to higher body

---

[16] The cdf is $H(EB \mid L, X_o, X_u; \xi, \mu) = \frac{\Gamma(\xi)}{\Gamma(\xi\mu)\Gamma(\xi(1-\mu))} \int_0^{EB} t^{\xi\mu-1} (1-t)^{\xi(1-\mu)-1} dt$. In Stata/Mata, EB can be generated by using the "invibeta(a,b,p)" function, which is the inverse beta cdf where a and b are the shape parameters, and p is a value between 0 and 1. In our case, a = ξμ, b = ξ(1 – μ), and p is the uniform random variable on the unit interval, i.e. U[0, 1].

[17] For a detailed discussion of the sampling design see Appendix 2B.

[18] An active female aged 14-18 years may need 2400 calories per day. For daily calorie needs for other age-gender groups at different levels of physical activity, see http://www.webmd.com/diet/features/estimated-calorie-requirement

fat %). By setting β_u at a positive value we assume that the unobservable component (X_u) is dominated by factors such as one's genetic predisposition to consume sugar which is positively related to both the consumption of calories from sugared soft drinks and body fat % (Qi et al., 2012).

Table 2.1.A and 2.1.B displays the summary statistics of the main variables in the simulated sample and the values of the key parameters (the αs and the βs.) chosen for the sampling design, respectively. It is clear to see that sample means of the four variables are all as expected. About 89% of the observations in the sample have their body fat % fall within 6% and 45%, which is quite reasonable. And the maximum amount of soda calories intake per day is about 461 cal., which is less than the amount of calories needed per day for an active male aged 14-18 years (3200 cal.). By design, the coefficient of X_u is nonzero ($\beta_u = 0.005$), meaning that soda calorie consumption variable is endogenous in the child energy balance model. The severity of the endogeneity problem is determined by the magnitude of β_u, i.e. the higher the magnitude of β_u, the more serious the problem will be.

2.4.2 Coefficient Parameter Estimation in the Simulated Simplified Model

Using the simulated data we estimated the parameters of the model in (2.1) and (2.3) via the 2SRI protocol given in section 2.1 [culminating in the NLS estimation of (2.4)]. For comparison, we also provide the relevant estimates obtained from: 1) a simple nonlinear regression (NR) method that ignores endogeneity (but not nonlinearity); and 2) the linear instrumental variables (LIV) method that ignores nonlinearity (but not endogeneity). For simplicity, the three types of coefficient estimates will be referred to as the 2SRI estimators, the NR estimators, and the LIV estimators, respectively.

The following two-stage protocol was used for the NR method,

Stage 1 – is the same as the one in the 2SRI method, i.e. estimate the lifestyle regression (2.1) using the NLS method, and thereby obtain consistent estimates of the αs (say $\tilde{\alpha}_o^{NR}$, $\tilde{\alpha}_P^{NR}$);

Stage 2 – estimate the energy balance regression (2.3) without including X_u by applying the NLS method, and obtain the corresponding estimates of the βs (say $\tilde{\beta}_L^{NR}$, $\tilde{\beta}_o^{NR}$).

Compared with the 2SRI method, in the NR approach the unobserved confounders for L, X_u, are not included in the energy balance regression, while the lifestyle regression is exactly the same in both methods. Therefore, the NR estimators for the βs based on the simulated data will not be consistent, but estimators for the αs will be consistent and equal to the 2SRI estimators. the LIV method

The LIV method is actually the two-stage least squares (2SLS) method, in which both two stages are based on linear regressions. The two stages are as follows:

Stage 1 – estimate the linearized lifestyle regression, $L = X_o\alpha_o + P\alpha_P + u$, by OLS, and obtain estimators, $\tilde{\alpha}_o^{LIV}$ and $\tilde{\alpha}_P^{LIV}$. Then construct $\tilde{L}^{LIV} = X_o\tilde{\alpha}_o^{LIV} + P\tilde{\alpha}_P^{LIV}$;

Stage 2 – estimate the linearized energy balance regression, $EB = \tilde{L}^{LIV}\beta_L + X_o\beta_o + e^{LIV}$, by OLS, and obtain estimates of the βs, $\tilde{\beta}_L^{LIV}$ and $\tilde{\beta}_o^{LIV}$.

The LIV method accounts for endogeneity but ignores inherent nonlinearity.

The first column of Table 2.2 shows the true parameter values (i.e. pre-specified values listed in Table 2.1.B) and the corresponding estimates obtained from each of the three methods. The 2SRI estimates listed in the second column are quite close to the true parameters, while the NR and the LIV estimates listed in the last two columns are far from

the true values; except, of course, for the NR estimates of the αs. Estimates of the αs obtained from the NR method are equal to the ones obtained from the 2SRI method, which is just as expected since the first stages in both methods are exactly the same. There is an upward bias for the NR coefficient estimate of soda calorie consumption ($\tilde{\beta}_L^{NR} = 0.00861$ vs $\beta_L = 0.007$), which is consistent with the fact that the unobserved confounders are assumed to be positively related to both soda consumption and body fat % in the simulated data. Although the LIV method accounts for the endogeneity of soda consumption, it incorrectly assumes linear functional forms for both the lifestyle regression and the energy balance regression, and hence produces inconsistent coefficient estimates. This is similar to the results obtained by Terza, Bradford and Dismuke (2008). Overall, the 2SRI estimator outperforms the other estimation approaches.

2.4.3 Average Incremental Effect Estimation in the Simulated Simplified Model

Using the simulated data and each of the three sets of parameter estimates (2SRI, NR, and LIV) we estimated the AIEs of an increment in soda calorie consumption (Δ_L) or soda price (Δ_P) on body fat %. The 2SRI-based estimates were obtained through direct application of (2.5) and (2.6) and are listed in the second columns of Tables 2.3.A and Table 2.3.B. For the NR case, the estimated AIEs were calculated via the versions of equations (2.5) and (2.6) that exclude the $\tilde{X}_{ui}^{2SRI}\tilde{\beta}_u^{2SRI}$ component and replace the $\tilde{\alpha}^{2SRI}$s and $\tilde{\beta}^{2SRI}$s by $\tilde{\alpha}^{NR}$s and $\tilde{\beta}^{NR}$s . These estimates are listed in the third columns of Tables 2.3.A and Table 2.3.B.[19] The LIV-estimated AIEs analogous to (2.5) and (2.6) are: $\Delta_L\tilde{\beta}_L^{LIV}$

[19] For detailed derivations of the NR-estimated AIEs see Appendix 2C.

and $\Delta_P\tilde{\alpha}_P^{LIV}\tilde{\beta}_L^{LIV}$, respectively, with $\tilde{\alpha}_P^{LIV}$ and $\tilde{\beta}_L^{LIV}$, and are listed in the fourth columns of Tables 2.3.A and Table 2.3.B.

In Table 2.3.A the simulation results for the 2SRI-, NR-, and, LIV-estimated AIE of a one calorie increment in soda consumption on body fat %, denoted by $\widetilde{AIE(\Delta_L)}^{2SRI}$, $\widetilde{AIE(\Delta_L)}^{NR}$, and $\widetilde{AIE(\Delta_L)}^{LIV}$ respectively. These three estimated AIEs are compared with the true value, denoted by $AIE(\Delta_L)$, which is calculated by substituting the true parameters (listed in the first column of Table 2.2) for the 2SRI estimators ($\tilde{\alpha}^{2SRI}$s and $\tilde{\beta}^{2SRI}$s) in equation (2.5).[20] The true value, $AIE(\Delta_L)$, is 0.001084, meaning that one more calorie from soda consumption will increase body fat % by around 0.11 of a percentage point on average. The corresponding estimate based on the 2SRI method is quite close to this value (0.001086 vs 0.001084), while the estimates obtained from the NR method and the LIV method are quite divergent from the true value (0.001338 vs 0.001084, and 0.001354 vs 0.001084, respectively).

The results are similar for the estimated AIE of a one dollar increment in soda price on body fat %. As can be seen in Table 2.3.B, the true value [$AIE(\Delta_P)$] is equal to -0.10503, indicating that a one dollar increase in soda price per 2 liters will decrease body fat % by around 10.5 percentage points. The NR- and LIV-estimated AIEs differ a lot from the true value

(-0.12633 vs -0.10503, and -0.30202 vs -0.10503, respectively), while the 2SRI-estimated AIE is very close to the population value (-0.10539 vs -0.10503).

[20] The true value of $AIE(\Delta_L)$ was calculated based on a super sample of 3 million observations generated using the same sampling design as that used to simulate the analysis sample of size 200,000. The true value for $AIE(\Delta_P)$ was similarly obtained; as were the true values Δ_L^0 and Δ_P^0 discussed in the following section.

We also test the performance of the 2SRI-estimated AIEs under different levels of endogeneity by adjusting the magnitude of β_u when generating data. We would like to see whether the AIE estimators based on our preferred 2SRI method are consistently superior to the ones based on the NR or the LIV method in all cases. We therefore simulate 5 slightly different samples of the same size (200,000) by increasing β_u from 0 to 0.02, with 0.005 as the increment, while keeping other parameters unchanged during data generating process. As β_u increases, the endogeneity problem gets worse. For each sample, we apply all three methods for estimating the AIEs, and then calculate the absolute percentage bias of each relative to the true value (absolute %bias = [|(estimated value - true value)/true value|]*100%) for the estimated ones**.** Table 2.4.A (2.4.B) shows the AIEs of a one unit increase in soda calorie consumption (soda price) on body fat % for increasing levels of endogeneity. The 2SRI-estimated AIEs and the NR-estimated AIEs are nearly identical when there is no endogeneity problem ($\beta_u = 0$).[21] However, as β_u increases, the increase in the percentage bias for the NR-based AIE estimators is striking: it increases from 0.02% to 85.51% for changes in soda calorie consumption (Table 2.4.A); and from 0.12% to 78.09% for a one dollar increase in soda price (Table 2.4.B). On the other hand, the percentage bias for the 2SRI-based AIE estimators is small in all the cases (always less than 1%). This indicates the importance of taking care of endogeneity when there is strong belief in the existence of unobservable confounders. Moreover, merely account for nonlinearity is not enough. The LIV-estimated AIEs are also subject to large bias especially when β_u is large,

[21] The differences are so minor that they disappear even if the numbers are rounded to six decimal places in Table 2.4.A and five places in Table 2.4.B. You can see the differences when the numbers are rounded to one more place, which are not shown in Table 2.4.A and 2.4.B: 0.0011129 vs 0.0011130 for a one unit change in calorie consumption; and -0.105938 vs -0.105942 for a one dollar change in soda price.

which indicates that the adverse effects of ignoring nonlinearity get worse when endogeneity is more prevalent, even if the method itself accounts for endogeneity.

Based on the results from Table 2.3.A (B) and Table 2.4.A (B), we may conclude that, compared with the NR- and the LIV-estimated AIEs, the estimates obtained via the 2SRI method are more reliable for assessing the effects of potential policy interventions, especially when the endogeneity problem is prevalent.

2.4.4 Policy Recommendations in the Simulated Simple Model

Using the simulated data and each of the three sets of parameter estimates (2SRI, NR, and LIV) we estimated policy recommendations [PR] (Δ_L^0 and Δ_P^0) for a given energy balance target (EB^0). For the present simulation study we specify the energy balance target to be a population average body fat % of 15%. This choice is motivated by the fact that the recommended healthy body fat percentages are 9-15% for boys and 14-21% for girls according to an online article.[22] The 2SRI-based PR estimates were obtained through direct application of (2.7) and (2.8) and are listed in the second columns of Tables 2.5.A and Table 2.5.B. For the NR case, the estimated PRs were calculated by using the versions of equations (2.7) and (2.8) that exclude the $\tilde{X}_{ui}^{2SRI}\tilde{\beta}_u^{2SRI}$ component and replace the $\tilde{\alpha}^{2SRI}$s and $\tilde{\beta}^{2SRI}$s by $\tilde{\alpha}^{NR}$s and $\tilde{\beta}^{NR}$s These estimates are listed in the third columns of Tables 2.5.A and Table 2.5.B.[23]

[22] The online article regarding the recommended healthy body fat % for boys and girls is available at: http://www.livestrong.com/article/194320-body-fat-percentage-for-children/. Information regarding children and adolescents' healthy body fat % by age and gender can be found via the following link: http://www.doctoragostini.com/childhoodobesity/id4.html.

[23] For detailed derivations of the NR-estimated AIEs see Appendix 2C.

Compared with the 2SRI- and the NR-based estimated changes in L or P required to achieve EB^0, those based on the LIV parameter estimates are relatively simple. The LIV-based $\tilde{\Delta}_L^0$ is the one that solves the equation $EB^0 = \sum_{i=1}^{n} \frac{1}{n}\left\{(L_i + \tilde{\Delta}_L^0)\tilde{\beta}_L^{LIV} + X_{oi}\tilde{\beta}_o^{LIV}\right\}$, and the LIV-based $\tilde{\Delta}_P^0$ is the one that solves the equation $EB^0 = \sum_{i=1}^{n} \frac{1}{n}\left\{(P_i + \tilde{\Delta}_P^0)\tilde{\alpha}_P^{LIV}\tilde{\beta}_L^{LIV} + (\tilde{\alpha}_o^{LIV}\tilde{\beta}_L^{LIV} + \tilde{\beta}_o^{LIV})X_{oi}\right\}$. Unlike equation (2.7) and (2.8), there are closed form solutions for $\tilde{\Delta}_L^0$ and $\tilde{\Delta}_P^0$, i.e. the LIV-based $\tilde{\Delta}_L^0$ and $\tilde{\Delta}_P^0$ are

$$\frac{EB^0 - \sum_{i=1}^{n} \frac{1}{n}\left\{L_i\tilde{\beta}_L^{LIV} + X_{oi}\tilde{\beta}_o^{LIV}\right\}}{\tilde{\beta}_L^{LIV}} \tag{2.15}$$

and

$$\frac{EB^0 - \sum_{i=1}^{n} \frac{1}{n}\left\{P_i\tilde{\alpha}_P^{LIV}\tilde{\beta}_L^{LIV} + (\tilde{\alpha}_o^{LIV}\tilde{\beta}_L^{LIV} + \tilde{\beta}_o^{LIV})X_{oi}\right\}}{\tilde{\alpha}_P^{LIV}\tilde{\beta}_L^{LIV}} \tag{2.16}$$

respectively. The LIV-estimated PRs, obtained using (2.15) and (2.16) are are listed in the fourth columns of Tables 2.5.A and Table 2.5.B.

We also estimated the PRs using the conventional approach prefer the approach (see section 2.3). Recall that in this approach, the PR estimates are obtained by simply dividing the difference between the targeted average body fat % (EB^0) and the current average body fat % ($\overline{EB}$) by the estimated AIE with the relevant causal increment set equal to 1 (e.g. using the 2SRI-based parameter estimates and AIE result $\tilde{\Delta}_L^0 = (0.15 - 0.23)/0.001086 \approx -73.66$, where 0.001086 is obtained from the second column of Table 2.3.A). There is no difference between these two approaches in obtaining

estimated required changes in L or P if the regressions used in the method have linear functional forms, e.g. as in the discussion of the LIV method.[24] In nonlinear models, however, the 2SRI and NR based estimates diverge. This is demonstrated in the sixth and seventh columns of Tables 2.5A and 2.5.B.

The "true" PR values, Δ_L^0 and Δ_P^0, and their linear approximations are denoted by "True" Δ_L^0 and "True" Δ_P^0 in Tables 2.5.A and 2.5.B. These true values, along with the 2SRI- NR- and LIV-based PR estimates indicate that the decrease in soda consumption necessary to bring the current average body fat % (23%) down to 15% is around 85 calories, and the increase in soda price required is around 48 cents. The 2SRI-based $\tilde{\Delta}_L^0$ and $\tilde{\Delta}_P^0$ obtained from our preferred approach are quite close to the true values (-85.40 vs -85.48, and 0.479 vs 0.482, respectively), while the corresponding NR- and the LIV-based estimates diverge from the true values quite substantially. Comparing the true Δ_L^0 with its linear approximation, the former

(-85.48) is larger than the latter (-73.87) in absolute value, which is consistent with the conclusion from Figure 2.1.A, i.e. the conventional approach tends to underestimate the decrease in L necessary to achieve EB^0. Similarly, the true Δ_P^0 (0.482) is smaller than its linear approximation (0.762), which is consistent with the illustration in Figure 2.1.B. Finally, as expected, the LIV-based $\tilde{\Delta}_L^0$ ($\tilde{\Delta}_P^0$) obtained from our approach is exactly the

[24] In equation (2.15), $\sum_{i=1}^{n} \frac{1}{n} \left\{ L_i \tilde{\beta}_L^{LIV} + X_{oi} \tilde{\beta}_o^{LIV} \right\}$ is actually the sample average for body fat %, $\overline{EB}$, while $\tilde{\beta}_L^{LIV}$ is the LIV-estimated AIE of one unit increment in L on EB. Therefore, equation (2.15) is identical to the one used in the conventional approach on obtaining the requisite change in L. Similarly for equation (2.16), in which $\sum_{i=1}^{n} \frac{1}{n} \left\{ P_i \tilde{\alpha}_P^{LIV} \tilde{\beta}_L^{LIV} + (\tilde{\alpha}_o^{LIV} \tilde{\beta}_L^{LIV} + \tilde{\beta}_o^{LIV}) X_{oi} \right\}$ is the sample average for body fat %, $\overline{EB}$, and $\tilde{\alpha}_P^{LIV} \tilde{\beta}_L^{LIV}$ is the LIV-estimated AIE of one unit increment in P on EB.

same as the one obtained from the conventional approach as the regressions used in the LIV method are linear.

We also compare the policy recommendation estimators ($\tilde{\Delta}_L^0$ and $\tilde{\Delta}_P^0$) based on our approach for different levels of endogeneity. The results are shown in Table 2.6.A and 2.6.B.[25] The absolute percentage bias for the 2SRI-based estimates is always small irrespective of the level of endogeneity, while the bias for the NR- or the LIV-based estimates increases as β_u increases. Therefore, our 2SRI-based approach to the estimation of requisite policy changes for achieving a specified energy balance target estimates is preferred especially when soda calorie consumption is severely endogenous.

2.5 Summary

In this chapter, we illustrate our method based on a simple case where only one energy balance variable, one policy variable, one lifestyle variable, and one observable control variable are involved, and simulate data to examine our proposed method in the estimation of 1) AIEs on body fat % in response to an exogenous change in soda calories intake or soda price and 2) quantitative policy recommendations for changing soda calories intake or soda price aimed at achieving an ideal body fat %. We derive, and develop Stata® code for, those econometric estimators, and apply it using simulated data. The simulation results show that, overall, the 2SRI method performs very well: all the estimates of the coefficients, the AIEs, and the change in soda calories intake or soda price required to achieve the average ideal body fat % are quite close to the true values. And the estimators

[25] The results for the requisite change in L or P when β_u is 0.02 are not shown in the table as Stata/Mata Optimize procedure fails to converge in this case.

obtained from the method that ignores the inherent nonlinearity of the model (i.e. the LIV method) or the potential endogeneity of soda calorie consumption (i.e. the NR method) in the modeling of child energy balance, body fat %, are subject to substantial bias. Moreover, our approach in obtaining the two policy recommendation estimators is more appropriate than the conventional approach that relies on linear approximation.

Appendix 2A

Data Generating for L in Stata/Mata

As Stata/Mata does not provide inverse incomplete gamma function, we cannot generate L directly based on (2.11). But we can get it indirectly. The incomplete gamma function $\gamma(s, x)$ can be expressed as

$$\gamma(s,\ x) = \Gamma(s)\,G(s, x) \tag{2A-1}$$

where G(s, x) denotes the cdf of the simple Gamma random variable with parameter s. Using (2A-1), let

$$j = \gamma(d,\ k) = \Gamma(d)\,G(d, k) \tag{2A-2}$$

Solving (2A-2) for k yields

$$k = \gamma^{-1}(d,\ j) = G^{-1}\left(d, \frac{j}{\Gamma(d)}\right) \tag{2A-3}$$

where $G^{-1}(d, P)$ denotes the inverse cdf of the simple Gamma random variable. Combining (2A-3) with (2.3) yields[26]

$$L = a\ G^{-1}(\nu,\ U[0,1])^{\frac{1}{p}} \tag{2A-4}$$

[26] The Generalized Gamma Distribution (GGD) we assume for L is a three parameter based distribution, which is a special case of the four parameter GGD used in Tadikamalla, 1979, in which the location parameter is 0. We generate L indirectly through generating a standard gamma random variate, say X, with shape parameter v, and making the transformation $L = aX^{1/p}$. In other words, if L ~ GG (a, v, p), then $\left(\frac{L}{a}\right)^{p} \sim \text{Gamma}(v, 1)$.

where U[0, 1] denotes the uniform random number on the unit interval. We can now generate L using $a \times \text{rgamma}(\nu, 1)^{\frac{1}{p}}$, where rgamma(m, n) is the Stata command used to randomly generate a gamma variate with shape parameter m and scale parameter n.

Appendix 2B

Background and Motivation for the Chosen Sampling Design

According to the Bureau of Labor Statistics, monthly average prices for non-diet Cola per 2 liters ranged from $1.310 to $1.367 in 2009.[27] So as an approximation, the mean price of soft drinks per 2 liters in the simulated data was set at $1.35. Fletcher et al. (2010b) use the NHANES III data and the NHANES 1999-2006 data to study the effects of soft drink taxes on childhood obesity. Their sample consists of children and adolescents between the ages of 3 and 18, with mean age equal to 10.513 and a standard deviation equal to 0.043. Therefore, we generate our sample with a mean of age of approximately 10.5 as we are interested in the children and adolescents between the ages of 2 and 19. Moreover, the mean calories from soft drinks in the previous 24 hours in Fletcher et al. (2010b) are 115.247 cal., with standard deviation equal to 2.531. Lin et al. (2011) use the 1998-2007 National Consumer Panel data and the 2003-2006 NHANES data. Their sample consists of children aged 2-19 years. The average calorie intake from sugar-sweetened beverages (SSBs) (including regular soft drinks, sports and energy drinks, and fruit drinks) is 189 cal. for children from low income families and 195 cal. for children from high income families. An online report shows that the energy obtained from SSBs for individuals aged 2-19 was 155 calories a day in 2009/2010.[28] Given the above information, we chose the average soda calorie intake per day to be around 120 cal.

Based on NHANES 1999-2004 data, a national health statistics report (Ogden, C.L. et al., 2011) shows that the mean percentage body fat at age 8 was 28% for boys and 31%

[27] See http://data.bls.gov/timeseries/APU0000717114?data_tool=XGtable
[28] See http://www.foodnavigator-usa.com/Markets/Calories-from-sugar-sweetened-beverages-have-declined-steadily-since-1999-but-still-account-for-11-of-energy-intakes-in-teenage-boys

for girls respectively, and these numbers decrease to 23% for boys at age 19 and increase to 35% for girls at age 19. Since we distinguish neither between boys and girls nor between different ages in our simplified illustration, we set the average body fat % at around 23%.

Appendix 2C

Background and Motivation for AIE Estimators and Recommended Requisite Price Change

The regression in (2.1) is causal in the sense that X_o is *comprehensive with respect to P*, i.e. it comprises all the possible confounders for P and its own elements.[29] Similarly, (2.3) is causal because $X = [X_o \quad X_u]$ is comprehensive with respect to L. In other words, conditional on X_o (X), any differences in the mean of the observed value of L (EB) can be exclusively attributed to differences in the observed value of P (L).

We distinguish between the observable version of P and its exogenously mandated version, P^*. Correspondingly, we define the observable version of L and its potential outcome version (L_{P^*}) – the version of L that would obtain if the policy variable were mandated to be P^*. Likewise we define the observable version of EB and its potential outcome version ($EB_{L_{P^*}}$) – the version of EB that would obtain if the policy variable were mandated to be P^*. To define the effect of a policy driven exogenous (and counterfactual) change in P on energy balance, we focus on how $EB_{L_{P^*}}$ would change between: the pre-policy scenario in which the distribution of P is exogenously set at $P^* = P^{pre}$; and the post-policy scenario in which the distribution of P^{pre} is exogenously incremented by Δ_P, a fixed constant. Within this framework, the 2SRI-based policy effect of interest can be formally defined as

$$AIE^{2SRI}(\Delta_P) = E[EB_{L_{P^{pre}+\Delta P}}] - E[EB_{L_{P^{pre}}}] \qquad (2C\text{-}1)$$

[29] See Terza (2014) for an expanded discussion of the concepts used here.

To simplify the discussion (and in keeping with convention), we assume that the observable value of L for any individual in the population is the same as it would have been if the observable value of P were exogenously imposed rather than the product of individual choice. In other words, for every individual in the population ω we have that

$$L(\omega) = L_{P^{exog}(\omega)}(\omega) \tag{2C-2}$$

where P^{exog} denotes the random variable representing the observable distribution of P treated as if it were exogenously imposed. Similarly, we assume that the observable value of EB for any individual in the population is the same as it would have been if the observable value of L were replaced by $L_{P^{exog}(\omega)}(\omega)$, where $L_{P^{exog}}$ is potential outcome version of L for P^{exog}. In other words, for every individual in the population ω

$$EB(\omega) = EB_{L_{P^{exog}(\omega)}}(\omega) \tag{2C-3}$$

Based on (2.9) and (2C-3)

$$EB_{L_{P^{exog}}} = \Lambda(L_{P^{exog}}\beta_L + X_o\beta_o + X_u\beta_u) + e \tag{2C-4}$$

and based on (2.7) and (2C-2)

$$L_{P^{exog}} = \exp(X_o\alpha_o + P^{exog}\alpha_P) + X_u \tag{2C-5}$$

Extending (2C-4) and (2C-5) to *any* exogenously imposed version (distribution) of the policy variable (say P^*) we obtain

$$EB_{L_{P^*}} = \Lambda(L_{P^*}\beta_L + X_o\beta_o + X_u\beta_u) + e \tag{2C-6}$$

and based on (2.7) and (2C-2)

$$L_{P^*} = \exp(X_o\alpha_o + P^*\alpha_P) + X_u \tag{2C-7}$$

Using the law of iterated expectations, it follows from (2C-6) and (2C-7) that

$$\begin{aligned} E[EB_{L_{P^*}}] &= E[\Lambda(L_{P^*}\beta_L + X_o\beta_o + X_u\beta_u)] + E\Big[E[e \mid L_{P^*}, X_o, X_u]\Big] \\ &= E[\Lambda([\exp(X_o\alpha_o + P^*\alpha_P) + X_u]\beta_L + X_o\beta_o + X_u\beta_u)] \end{aligned} \tag{2C-8}$$

because, by assumption, $E[e \mid L_{P^*}, X_o, X_u] = 0$. For our analyses, we will follow the typical approach and take the hypothetically mandated pre-policy version of the policy variable to be $P^{pre} = P^{exog}$. Combining this assumption with (2C-1) and (2C-8) yields

$$AIE^{2SRI}(\Delta_P) = E[\Lambda([\exp(X_o\alpha_o + (P^{exog} + \Delta_P)\alpha_P) + X_u]\beta_L + X_o\beta_o + X_u\beta_u)] - E[EB] \tag{2C-9}$$

Clearly, the statistic given in (2.12) is the sample analog to (2C-9). It is, therefore, easy to show that since $\tilde{\alpha}_o^{2SRI}$, $\tilde{\alpha}_P^{2SRI}$, $\tilde{\beta}_L^{2SRI}$, $\tilde{\beta}_o^{2SRI}$ and $\tilde{\beta}_u^{2SRI}$ are consistent estimators of α_o, α_P, β_L, β_o and β_u, respectively, (2.12) is a consistent estimator of (2C-9).

If we do not know the exact change in a particular policy that impacts L, but have information about the resultant change in L between the pre-policy scenario and the post-

policy scenario, the effect of an exogenous policy-driven shift in L by the amount Δ_L on energy balance can be formally defined as

$$AIE^{2SRI}(\Delta_L) = E[EB_{L^{pre}+\Delta_L}] - E[EB_{L^{pre}}] \quad (2C\text{-}10)$$

where L^{pre} is the distribution of L in the pre-policy scenario. Assume that, for every individual in the population ω, we have that

$$EB(\omega) = EB_{L^{exog}(\omega)}(\omega) \quad (2C\text{-}11)$$

where L^{exog} denotes the random variable representing the observable distribution of L treated as if it were exogenously imposed. (2C-11) indicates that the observable value of EB for any individual in the population is the same as it would have been if the observable value of L were exogenously imposed rather than the product of individual choice. Based on (2.9) and (2C-11)

$$EB_{L^{exog}} = \Lambda(L^{exog}\beta_L + X_o\beta_o + X_u\beta_u) + e. \quad (2C\text{-}12)$$

Extending (2C-12) to any exogenously imposed version of L (say L^*, analogous to P^*) we obtain

$$EB_{L^*} = \Lambda(L^*\beta_L + X_o\beta_o + X_u\beta_u) + e. \quad (2C\text{-}13)$$

Using the law of iterated expectations, it follows from (2C-13) that

$$E[EB_{L^*}] = E[\Lambda(L^*\beta_L + X_o\beta_o + X_u\beta_u)] + E\Big[E[e \mid L^*, X_o, X_u]\Big]$$

$$= E[\Lambda(L^{*}\beta_L + X_o\beta_o + X_u\beta_u)] \quad \text{(2C-14)}$$

as $E[e \mid L^{*}, X_o, X_u] = 0$ by assumption. Assume $L^{exog} = L^{pre}$, and combine (2C-10) and (2C-14), we have

$$AIE^{2SRI}(\Delta_L) = E[\Lambda((L^{exog} + \Delta_L)\beta_L + X_o\beta_o + X_u\beta_u)] - E[EB]. \quad \text{(2C-15)}$$

Clearly, (2C-15) can be consistently estimated by (2.11).

By applying a similar approach, the effects of P or L on energy balance analogous to (2C-9) and (2C-15) but imposing the condition that L is exogenous are

$$AIE^{NR}(\Delta_P) = E[\Lambda([\exp(X_o\alpha_o + (P^{exog} + \Delta_P)\alpha_P) + X_u]\beta_L + X_o\beta_o)] - E[EB] \quad \text{(2C-16)}$$

and

$$AIE^{NR}(\Delta_L) = E[\Lambda((L^{exog} + \Delta_L)\beta_L + X_o\beta_o)] - E[EB] \quad \text{(2C-17)}$$

respectively, and (2C-16) and (2C-17) can be consistently estimated by

$$\left(\sum_{i=1}^{n}\frac{1}{n}\Lambda([\exp(X_{oi}\tilde{\alpha}_o^{NR} + (P_i + \Delta_P)\tilde{\alpha}_P^{NR}) + \tilde{X}_{ui}^{NR}]\tilde{\beta}_L^{NR} + X_{oi}\tilde{\beta}_o^{NR})\right) - \overline{EB} \quad \text{(2C-18)}$$

and

$$\left(\sum_{i=1}^{n}\frac{1}{n}\Lambda((L_i + \Delta_L)\tilde{\beta}_L^{NR} + X_{oi}\tilde{\beta}_L^{NR})\right) - \overline{EB} \quad \text{(2C-19)}$$

where $\tilde{X}_{ui}^{NR} = L_i - \exp(X_{oi}\tilde{\alpha}_o^{NR} + P_i\tilde{\alpha}_P^{NR})$. The LIV-based effects are relatively simple as both lifestyle regression and energy balance regression are linear so that all component that

remain unchanged from pre- to post-policy will cancel out. The two types of effect based on the LIV method analogous to (2C-9) and (2C-15), are

$$\mathrm{AIE}^{\mathrm{LIV}}(\Delta_{\mathrm{P}}) = \Delta_{\mathrm{P}}\alpha_{\mathrm{P}}\beta_{\mathrm{L}} \tag{2C-20}$$

and

$$\mathrm{AIE}^{\mathrm{LIV}}(\Delta_{\mathrm{L}}) = \Delta_{\mathrm{L}}\beta_{\mathrm{L}} \tag{2C-21}$$

respectively, and can be estimated as $\Delta_{\mathrm{P}}\tilde{\alpha}_{\mathrm{P}}^{\mathrm{LIV}}\tilde{\beta}_{\mathrm{L}}^{\mathrm{LIV}}$ and $\Delta_{\mathrm{L}}\tilde{\beta}_{\mathrm{L}}^{\mathrm{LIV}}$ correspondingly.

We now turn to characterize the policy-driven change in P (Δ_{P}^{0}) or L (Δ_{L}^{0}) that would be required to bring the average energy balance to a targeted level ($\mathrm{E}[\mathrm{EB}] = \mathrm{EB}^{0}$). As in the above discussion, we take $\mathrm{P}^{\mathrm{exog}}$ or $\mathrm{L}^{\mathrm{exog}}$ as the pre-policy starting point. Using (2C-8) we get

$$\mathrm{EB}^{0} = \mathrm{E}[\Lambda([\exp(\mathrm{X}_{\mathrm{o}}\alpha_{\mathrm{o}} + (\mathrm{P}^{\mathrm{exog}} + \Delta_{\mathrm{P}}^{0})\alpha_{\mathrm{P}}) + \mathrm{X}_{\mathrm{u}}]\beta_{\mathrm{L}} + \mathrm{X}_{\mathrm{o}}\beta_{\mathrm{o}} + \mathrm{X}_{\mathrm{u}}\beta_{\mathrm{u}})] \tag{2C-22}$$

and the requisite policy change is the value of Δ_{P}^{0} that solves (2C-22). Using (2C-14) we get

$$\mathrm{EB}^{0} = \mathrm{E}[\Lambda((\mathrm{L}^{\mathrm{exog}} + \Delta_{\mathrm{L}}^{0})\beta_{\mathrm{L}} + \mathrm{X}_{\mathrm{o}}\beta_{\mathrm{o}} + \mathrm{X}_{\mathrm{u}}\beta_{\mathrm{u}})] \tag{2C-23}$$

and the requisite policy-driven change in L is the value of Δ_{L}^{0} that solves (2C-23). The solution, with respect to $\tilde{\Delta}_{\mathrm{P}}^{0}$ ($\tilde{\Delta}_{\mathrm{L}}^{0}$), to the version of (2.14) [(2.13)] that replaces the $\tilde{\alpha}$s and $\tilde{\beta}$s by $\tilde{\alpha}^{\mathrm{2SRI}}$s and $\tilde{\beta}^{\mathrm{SRI}}$s is the sample analog to Δ_{P}^{0} (Δ_{L}^{0}) in (2C-22) [(2C-23)], and is the

2SRI-based estimated change in P (L) required to achieve EB^0. It is easy to show that 2SRI-based $\tilde{\Delta}_P^0$ ($\tilde{\Delta}_L^0$) is a consistent estimator of Δ_P^0 (Δ_L^0) as the $\tilde{\alpha}^{2SRI}$s and $\tilde{\beta}^{SRI}$s are consistent estimators.

Similarly, imposing the exogeneity of L, the requisite policy-driven change in P or L is the value of Δ_P^0 or Δ_L^0 that solves

$$EB^0 = E[\Lambda([\exp(X_o\alpha_o + (P^{exog} + \Delta_P^0)\alpha_P) + X_u]\beta_L + X_o\beta_o)] \qquad (2C\text{-}24)$$

or

$$EB^0 = E[\Lambda((L^{exog} + \Delta_L^0)\beta_L + X_o\beta_o)] \qquad (2C\text{-}25)$$

and the corresponding NR-based estimator $\tilde{\Delta}_P^0$ ($\tilde{\Delta}_L^0$) is the solution to the version of equation (2.14) [(2.13)] with respect to $\tilde{\Delta}_P^0$ ($\tilde{\Delta}_L^0$) that excludes the $\tilde{X}_{ui}\tilde{\beta}_u$ component and substitutes the $\tilde{\alpha}^{NR}$s and $\tilde{\beta}^{NR}$s for the $\tilde{\alpha}$s and $\tilde{\beta}$s.

Compared with the 2SRI- and the NR-based policy recommendation estimators defined in (2C-22) – (2C-25), those based on LIV parameter estimates are relatively simple. By applying a similar approach, they can be easily derived. We will not repeat the derivations here.

Table 2.1.A: Summary Statistics of the Simulated Sample in Simple Case

Variables	Definition	Mean	St. dev.
P	Soda price ($ per 2 liters)	1.35	0.32
X_o[30]	Age (years old)	10.5	4.9
L	Soda calorie consumption per day (cal.)	120.06	88.54
EB	Body fat %	0.23 (23%)	0.15 (15%)

Table 2.1.B: Pre-specified Values of the α and β Parameters

Parameters	α_o	α_P	β_L	β_o	β_u
Values	0.001	-2	0.007	0.02	0.005

[30] X_o is actually a vector consists of two elements, age and a scalar 1 (used for adding a constant in the equation). For simplicity, here we just use it to denote age. Similarly, we use α_o and β_o to denote the coefficients of age in lifestyle equation and energy balance equation respectively, i.e. they are scalars, not vectors.

Table 2.2: Coefficient Estimates

True Parameters	2SRI Estimators	NR Estimators	LIV Estimators
$\alpha_o = 0.001$	$\tilde{\alpha}_o^{2SRI} = 0.00109$	$\tilde{\alpha}_o^{NR} = 0.00109$	$\tilde{\alpha}_o^{LIV} = 0.12302$
$\alpha_P = -2$	$\tilde{\alpha}_P^{2SRI} = -2.00405$	$\tilde{\alpha}_P^{NR} = -2.00405$	$\tilde{\alpha}_P^{LIV} = -223.09$
$\beta_L = 0.007$	$\tilde{\beta}_L^{2SRI} = 0.00701$	$\tilde{\beta}_L^{NR} = 0.00861$	$\tilde{\beta}_L^{LIV} = 0.00135$
$\beta_o = 0.02$	$\tilde{\beta}_o^{2SRI} = 0.02007$	$\tilde{\beta}_o^{NR} = 0.01955$	$\tilde{\beta}_o^{LIV} = 0.00309$
$\beta_u = 0.005$	$\tilde{\beta}_u^{2SRI} = 0.00498$	-	-

Table 2.3.A: Average Incremental Effects (AIEs) of Daily Soda Calorie Consumption (L) on Child's Body Fat % (EB), with 1 Calorie Increment in L ($\Delta L = 1$)

True $AIE(\Delta_L)$	$\widetilde{AIE(\Delta_L)}^{2SRI}$	$\widetilde{AIE(\Delta_L)}^{NR}$	$\widetilde{AIE(\Delta_L)}^{LIV}$
0.001084	0.001086	0.001338	0.001354

Table 2.3.B: Average Incremental Effects (AIEs) of Soda Price (P) on Child's Body Fat % (EB), with 1 Dollar Increment in P ($\Delta_P = 1$)

True $AIE(\Delta_P)$	$\widetilde{AIE(\Delta_P)}^{2SRI}$	$\widetilde{AIE(\Delta_P)}^{NR}$	$\widetilde{AIE(\Delta_P)}^{LIV}$
-0.10503	-0.10539	-0.12633	-0.30202

Table 2.4.A: Average Incremental Effects (AIEs) of Daily Soda Calorie Consumption (L) on Child's Body Fat % (EB), with 1 Calorie Increment in L (ΔL = 1), for Increasing Level of Endogeneity

Level of Endogeneity (β_u)	True $AIE(\Delta_L)$	$\widetilde{AIE(\Delta_L)}^{2SRI}$	$\widetilde{AIE(\Delta_L)}^{NR}$	$\widetilde{AIE(\Delta_L)}^{LIV}$	Absolute %bias (2SRI)	Absolute %bias (NR)	Absolute %bias (LIV)
0	0.001113	0.001113	0.001113	0.001268	0.03%	0.02%	13.92%
0.005	0.001084	0.001086	0.001338	0.001354	0.24%	23.46%	24.94%
0.01	0.001050	0.001055	0.001528	0.001455	0.50%	45.52%	38.54%
0.015	0.001015	0.001023	0.001687	0.001553	0.75%	66.13%	52.90%
0.02	0.000982	0.000992	0.001821	0.001639	1.00%	85.51%	66.92%

Table 2.4.B: Average Incremental Effects (AIEs) of Soda Price (P) on Child's Body Fat % (EB), with 1 Dollar Increment in P (ΔP = 1), for Increasing Level of Endogeneity

Level of Endogeneity (β_u)	True $AIE(\Delta_P)$	$\widetilde{AIE(\Delta_P)}^{2SRI}$	$\widetilde{AIE(\Delta_P)}^{NR}$	$\widetilde{AIE(\Delta_P)}^{LIV}$	Absolute %bias (2SRI)	Absolute %bias (NR)	Absolute %bias (LIV)
0	-0.10582	-0.10594	-0.10594	-0.28292	0.12%	0.12%	167.38%
0.005	-0.10503	-0.10539	-0.12633	-0.30202	0.34%	20.28%	187.56%
0.01	-0.10324	-0.10382	-0.14389	-0.32449	0.56%	39.37%	214.29%
0.015	-0.10032	-0.10111	-0.15896	-0.34636	0.78%	58.45%	245.25%
0.02	-0.09656	-0.09754	-0.17197	-0.36560	1.01%	78.09%	278.61%

Table 2.5.A: Simulation Results for the Target Policy (Δ_L^0) to Achieve Average Ideal Body Fat % (EB0) =15% -- Comparison of Our Approachand the Conventional Approach

Our Approach				The Conventional Approach			
True Δ_L^0	2SRI $\tilde{\Delta}_L^0$	NR $\tilde{\Delta}_L^0$	LIV $\tilde{\Delta}_L^0$	"True" Δ_L^0	2SRI $\tilde{\Delta}_L^0$	NR $\tilde{\Delta}_L^0$	LIV $\tilde{\Delta}_L^0$
-85.48	-85.40	-68.46	-59.12	-73.87	-73.69	-59.83	-59.12

Table 2.5.B: Simulation Results for the Target Policy (Δ_P^0) to Achieve Average Ideal Body Fat % (EB^0) =15% -- Comparison of Our Approachand the Conventional Approach

Our Approach				The Conventional Approach			
True Δ_P^0	2SRI $\tilde{\Delta}_P^0$	NR $\tilde{\Delta}_P^0$	LIV $\tilde{\Delta}_P^0$	"True" Δ_P^0	2SRI $\tilde{\Delta}_P^0$	NR $\tilde{\Delta}_P^0$	LIV $\tilde{\Delta}_P^0$
0.482	0.479	0.322	0.265	0.762	0.759	0.634	0.265

Table 2.6.A: Simulation Results for the Target Policy (Δ_L^0) to Achieve Average Ideal Body Fat% (EB0) =15% for Increasing Level of Endogeneity, Using Our Approach

Level of Endogeneity (β_u)	Sample Mean of Body Fat %	True Δ_L^0	2SRI $\tilde{\Delta}_L^0$	NR $\tilde{\Delta}_L^0$	LIV $\tilde{\Delta}_L^0$	Absolute %bias (2SRI)	Absolute %bias (NR)	Absolute %bias (LIV)
0	0.22 (22%)	-74.69	-74.75	-74.77	-57.13	0.09%	0.12%	23.51%
0.005	0.23 (23%)	-85.48	-85.40	-68.46	-59.12	0.09%	19.91%	30.84%
0.01	0.24 (24%)	-100.11	-99.85	-67.32	-61.75	0.25%	32.75%	38.32%
0.015	0.25 (25%)	-117.60	-117.06	-68.74	-64.81	0.46%	41.55%	44.89%

Table 2.6.B: Simulation Results for the Target Policy (Δ_P^0) to Achieve Average Ideal Body Fat% (EB0) =15% for Increasing Level of Endogeneity, Using Our Approach

Level of Endogeneity (β_u)	Sample Mean of Body Fat %	True Δ_P^0	2SRI $\tilde{\Delta}_P^0$	NR $\tilde{\Delta}_P^0$	LIV $\tilde{\Delta}_P^0$	Absolute %bias (2SRI)	Absolute %bias (NR)	Absolute %bias (LIV)
0	0.22 (22%)	0.384	0.384	0.384	0.256	0.11%	0.13%	33.33%
0.005	0.23 (23%)	0.482	0.479	0.322	0.265	0.46%	33.18%	44.97%
0.01	0.24 (24%)	0.655	0.649	0.304	0.277	0.89%	53.58%	57.76%
0.015	0.25 (25%)	1.000	0.983	0.306	0.291	1.69%	69.37%	70.96%

Figure 2.1.A

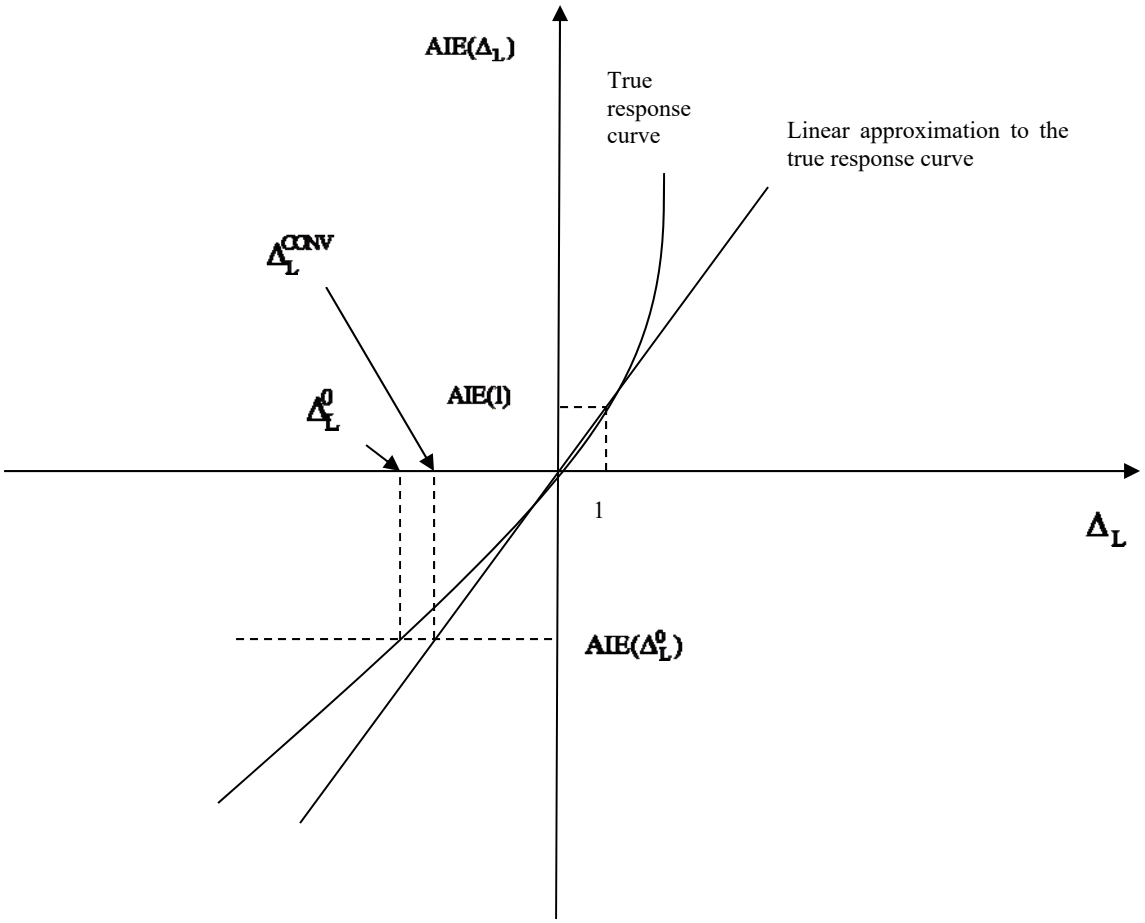
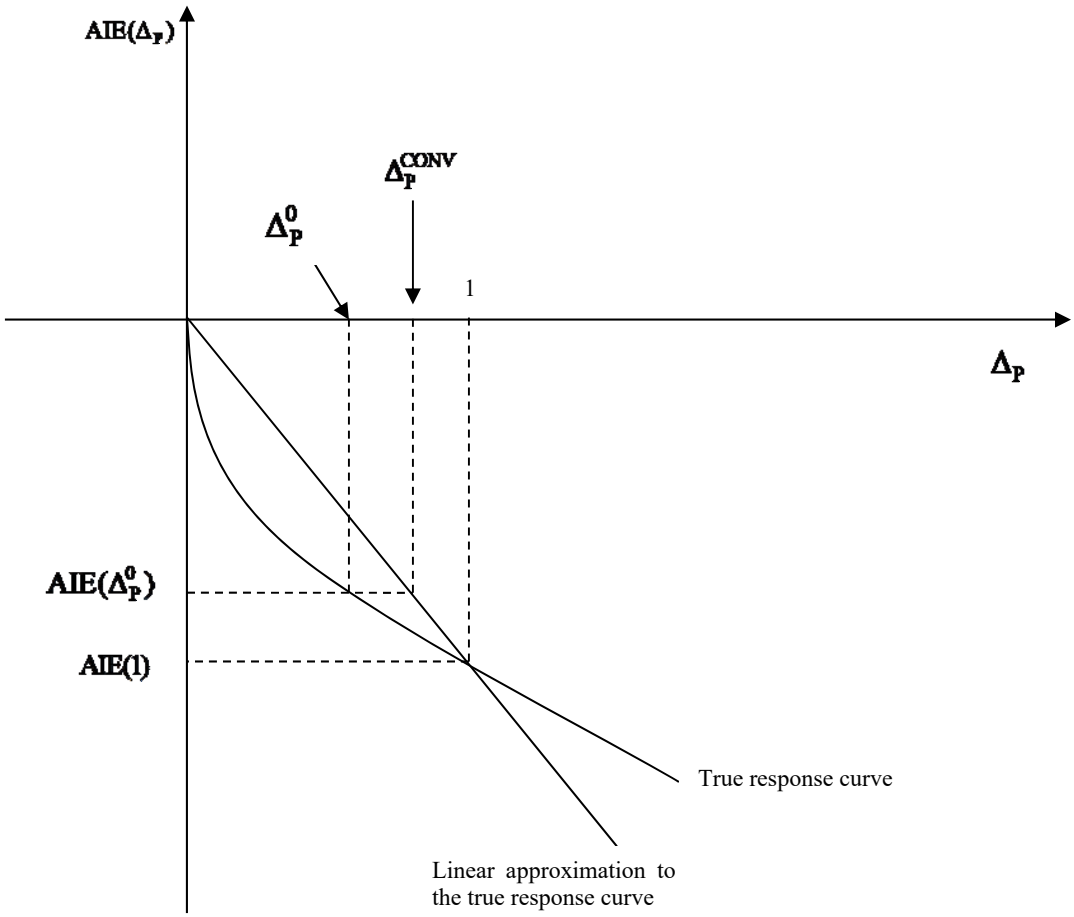

AIE(Δ_L)
True response curve
Linear approximation to the true response curve
Δ_L^{CONV}
Δ_L^0
AIE(1)
1
Δ_L
AIE(Δ_L^0)

Figure 2.1.B


AIE(Δ_P)
Δ_P^{CONV}
Δ_P^0
1
Δ_P
AIE(Δ_P^0)
AIE(1)
True response curve
Linear approximation to the true response curve

Chapter 3: The General Model and Proposed Empirical Policy Analytic Methods

In chapter 2, we examined a simplified version of the model by using simulated data, and showed that all the relevant estimators derived based on this model outperform those based on conventional models that ignore nonlinearity or endogeneity. This simple model provides a good illustration of our methods in obtaining causally-interpretable evidence on the impact of SSB consumption on children's weight and quantitative recommendations for potential SSB-related interventions aimed at specified energy balance goals. However, the situation described by the model is too simple: only one measure for each type of variable, i.e. energy balance variable, lifestyle variable, policy variable, observed control variable, is far from reality. Therefore, to make the model more general and more representative of the real world, in this chapter, we extend the simple univariate model to accommodate multiple policy and lifestyle variables. The reasons are quite obvious: many lifestyle choices may affect one's body weight, and each lifestyle choice variable is likely to depend on a number of potential policy levers. We also consider different measures of the energy balance outcome, as no single measure has been proven to be the most accurate indicator for obesity. We complete the discussion with an empirical application, showing the way of implementing our method in practice. Correct asymptotic standard errors of the relevant estimators are derived and coded in Stata®.

3.1 Regression Representation of the General Model

The lifestyle regression and the energy balance regression, analogous to (2.1) and (2.3), in the general model are

$$L_j = g_j\left(\underline{X}_o\alpha_{oj} + \underline{P}\alpha_{Pj}\right) + X_{uj} \tag{3.1}$$

and

$$EB_r = f_r\left(\underline{L}\beta_{Lr} + \underline{X}_o\beta_{or} + \underline{X}_u\beta_{ur}\right) + e_r \qquad (3.2)$$

respectively, where $g_j(\ )$ and $f_r(\ )$ are known functions ($r = 1, ..., R$; $j = 1, ..., J$); $\underline{L} = [L_1 \ldots L_J]$ is the vector comprising multiple lifestyle choice variables; $\underline{X}_o$ is a vector of regression control variables; $\underline{X}_u = [X_{u1} \ldots X_{uJ}]$ is the vector of unobserved confounders for the lifestyle choice variables in $\underline{L}$; $\underline{P} = [P_1 \ldots P_K]$ is the vector comprising a variety of potential policy variables ($k = 1, ..., K$); the αs and βs are regression parameters to be estimated, and e_r is the regression error term for the rth energy balance regression. This model is causal in the sense that it explicitly accounts for the potential endogeneity of lifestyle choice variables by including unobserved confounders $\underline{X}_u$ for $\underline{L}$ in (3.2) and all the other variables, $\underline{X}_o$ and $\underline{P}$, are assumed to be exogenous in both (3.1) and (3.2). The αs and βs can be consistently estimated by applying the 2SRI instrumental variables method, where the instruments are the policy variables included in $\underline{P}$ and are assumed to be highly correlated with lifestyle choice variables, but only correlated with energy balance outcome through their impacts on lifestyle choices. In order to identify the model, the number of policy variables (K) should be no less than the number of endogeneous variables (J), i.e. $K \geq J$.

To make the model more straightforward, let's focus on three lifestyle choices — SSB calories intake (L_1), other calories intake (L_2), minutes of physical activity per day (L_3); and four energy balance measures — BMI percentile (EB_1), body fat % (EB_2),

indicator for overweight (EB_3), indicator for obesity (EB_4). Then the lifestyle choice can be modeled via a nonnegative nonlinear regression of the form[31]

$$L_j = \exp(\underline{X}_o \alpha_{oj} + \underline{P}\alpha_{pj}) + X_{uj} \tag{3.3}$$

where j = 1, 2, 3; $\underline{X}_o$ and $\underline{P}$ are defined as in (3.1); the αs are the parameters to be estimated; and X_{uj} denotes the random error term. The energy balance outcome can be modeled as[32]

$$EB_r = \Lambda(\underline{L}\beta_{Lr} + \underline{X}_o\beta_{or} + \underline{X}_u\beta_{ur}) + e_r \tag{3.4}$$

where r = 1, 2, 3, 4; Λ( ) is the logistic cumulative distribution function (cdf); $\underline{L} = [L_1 \quad L_2 \quad L_3]$; $\underline{X}_o$ is defined as in (3.2); $\underline{X}_u = [X_{u1} \quad X_{u2} \quad X_{u3}]$; the βs are the parameters to be estimated; e_r is the random error term. To identify this model, we need the number of policy variables (K) to be no less than 3, i.e. $K \geq 3$. The 2SRI estimator for this model is:

Stage 1 – estimate each lifestyle equation in (3.3) via the nonlinear least squares (NLS) method, and obtain consistent coefficient estimates $\tilde{\alpha}_{oj}$, $\tilde{\alpha}_{Pj}$. Then construct the vector $\underline{\tilde{X}}_u = \left[\tilde{X}_{u1} \quad \tilde{X}_{u2} \quad \tilde{X}_{u3}\right]$, where $\tilde{X}_{uj} = L_j - \exp(\underline{X}_o \tilde{\alpha}_{oj} + \underline{P}\tilde{\alpha}_{pj})$.

Stage 2 – obtain consistent coefficient estimates, $\tilde{\beta}_{Lr}$, $\tilde{\beta}_{or}$, $\tilde{\beta}_{ur}$, by applying the NLS method to the following version of (3.4)

[31] Assuming lifestyle variables, i.e. L_1, L_2, and L_2 to be Generalized Gamma distribution.
[32] If using beta regressions for EB_1 and EB_2, and logit regressions for EB_3 and EB_4, (3.4) can be used to model all the four energy balance variables.

$$EB_r = \Lambda(\underline{L}\beta_{Lr} + \underline{X}_o\beta_{or} + \underline{\tilde{X}}_u\beta_{ur}) + e_r^{2SRI} \quad (3.5)$$

where $\underline{\tilde{X}}_u$ has been obtained from stage 1. These two stages would be repeated for each of the four energy balance outcomes.

3.2 Average Incremental Effects in the General Model

Based on the above model specification and the 2SRI coefficient estimates, we can now construct estimators for the lifestyle choice effects or the policy effects on a particular energy balance outcome. The average incremental effect (AIE), analogous to (2.5), of an exogenous policy-driven shift in the jth lifestyle variable by the amount δ_j on a specified energy balance measure can be estimated as

$$\left(\sum_{i=1}^{n}\frac{1}{n}\widetilde{EB_{ri}(\delta_j)}\right) - \overline{EB_r} \quad (3.6)$$

where $\overline{EB_r} = \sum_{i=1}^{n}\frac{1}{n}EB_{ri}$, EB_{ri} is the observed level of the rth energy balance measure for the ith individual in a sample of size n; $\widetilde{EB_{ri}(\delta_j)} = \Lambda\left(\underline{\mathcal{L}\left(L_{ji}+\delta_j\right)}\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)$, $\underline{\mathcal{L}\left(L_{ji}+\delta_j\right)}$ is the same as $\underline{L}_i$ with its jth element shifted by δ_j, $\underline{\tilde{X}}_{ui} = \begin{bmatrix}\tilde{X}_{u1i} & \tilde{X}_{u2i} & \tilde{X}_{u3i}\end{bmatrix}$, $\tilde{X}_{uji} = L_{ji} - \exp(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{P}_i\tilde{\alpha}_{pj})$; the $\tilde{\alpha}$s and the $\tilde{\beta}$s are the 2SRI coefficient estimates. Similarly, the estimated AIE, analogous to (2.6), of an exogenous change in the kth policy variable, say Δ_k, on a specified energy balance measure is

$$\left(\sum_{i=1}^{n}\frac{1}{n}\widetilde{EB_{ri}(\Delta_k)}\right) - \overline{EB_r} \quad (3.7)$$

where $\widetilde{EB_{ri}(\Delta_k)} = \Lambda\left(\underline{\tilde{L}_i(\Delta_k)}\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)$,

$\underline{\tilde{L}_i(\Delta_k)} = \left[\tilde{L}_{1i}(\Delta_k) \quad \tilde{L}_{2i}(\Delta_k) \quad \tilde{L}_{3i}(\Delta_k)\right]$, $\tilde{L}_{ji}(\Delta_k) = \exp(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{\boldsymbol{\mathcal{P}}(P_{ki}+\Delta_k)}\tilde{\alpha}_{Pj}) + \tilde{X}_{uji}$,

$\tilde{X}_{uji}$ is defined as in (3.6), $\underline{\boldsymbol{\mathcal{P}}(P_{ki}+\Delta_k)}$ is the same as $\underline{P}$ with its kth element replaced by $P_{ki}+\Delta_k$; $\overline{EB_r}$ denotes the sample average for the rth energy balance measure, defined as in (3.6).

It is easy to show that the two types of estimated AIEs, (3.6) and (3.7), are consistent as the parameter estimators, the $\tilde{\alpha}$s and $\tilde{\beta}$s, used in both equations are consistent. The lifestyle effect estimator in (3.6) can be used to evaluate the direct effects of exogenous changes in the lifestyle variables (however motivated) on children's energy balance, and compare the effectiveness of less formal (and possibly more direct) efforts to change children's behavior (i.e. less formal than policy measures based on manipulation of the elements of $\underline{P}$). It can also answer the question as to whether calories from SSBs differ from the same amount of calories from other sources in terms of affecting one's energy balance. The policy effect estimator in (3.7) can be used to evaluate and compare extant and planned obesity related policies that are based on a particular energy balance measure (EB_r) and the individual policy lever variables in $\underline{P}$.

3.3 Empirical Application

Our ultimate goals are to study the causal impact of sugar-sweetened beverage (SSB) consumption on childhood obesity, and provide quantitative policy recommendations for prospective policy interventions aimed at specified energy balance

goals. These objectives will be fulfilled by using the empirical data from the National Health and Nutrition Examination Survey (NHANES), combined with county level policy data, such as prices of foods and drinks, access to fast-food restaurants, grocery stores, etc., obtained from various sources. These aggregate-level policy variables are assumed to be exogenous in both lifestyle regression equations and energy balance equations, and highly correlated with lifestyle variables, and, hence, can be used as instrumental variables for lifestyle variables. However, the use of those variables requires state and county identifiers of the subjects involved in each wave of the NHANES, which are only available in the restricted data available through Census Research Data Centers (RDC). We are currently in the process of getting those data, which, we believe, will be ready soon. For now, as an illustration of our method, and a preliminary test as well, we use part of the data to demonstrate the feasibility of our proposed method in practice. Specifically, we use school breakfast policy, family Food Stamps receiving status and family frequency of eating at restaurant per week as instrumental variables for SSB calories intake, other calories intake, and physical activities, and study the impacts of those lifestyle variables on children's body fat %. . These instrumental variables are publicly available but are likely to be of lower quality than the aggregate-level policy variables (e.g. prices) that we will be able to obtain through the RDC. In particular these publicly available IVs are likely to be weak and probably violate the requisite IV validity condition. For this reason, we view the following empirical analyses as mainly illustrative and confine our causal analyses to the estimation of the AIEs of the lifestyle variable on energy balance. We forego estimation of the AIEs of the policy variables [as in (3.7)] (and estimation of recommend policy changes as detailed in Chapter2). More complete, and we expect more policy relevant, analyses will

be conducted once we secure the restricted aggregated data via the RDC and merge it with the NHANES data.

In the following two sub-sections, we first describe the NHANES and the aggregate-level policy data in detail, i.e. covering all the relevant variables that will be involved in our more complete analyses based on the restricted data. We then conduct an empirical analysis to illustrate our method using the data that does not require the usage of state and county identifiers.

3.3.1 Data

We use all the seven waves of the National Health and Nutrition Examination Survey (NHANES) data from 1999 to 2012 to construct our analysis sample. The NHANES repeatedly collects data from a multistage probability sample of the US civilian noninstitutionalized population since 1999, and releases it in a two-year cycle, i.e. 1999-2000, 2001-2002, etc. It is designed to assess health and nutritional status of children and adults. The survey consists of a home interview, during which the information of participant's demographic characteristics and physical activities are collected,[33] followed by a standardized physical examination in a mobile examination center (MEC). The examination includes physical measurements such as standing height, body weight, percent body fat[34], etc. A 24-hour diet recall interview is also conducted in the MEC.[35]

[33] For those participants aged 12 -15, physical activity information is collected in the mobile examination center (MEC).
[34] Percent body fat is only available in three waves, i.e. 1999-2000, 2001-2002, and 2003-2004, while other body measures, such as body weight and standing height, are available for all the seven waves.
[35] Started from 2003, NHANES releases two days of dietary data, among which day 1 data is collected in the MEC while day 2 data is collected via a phone interview. For wave 1999-2000 and 2001-2002, only one day dietary data was released and it was collected in the MEC.

We pool all seven waves of the data and restrict our analysis sample to include children aged 2-19. We construct the BMI percentile variable by comparing children's BMI, defined as body weight divided by squared height (kg/m^2), to the 2000 Centers for Disease Control and Prevention (CDC) gender-specific BMI-for-age growth charts. Then the overweight indicator is set to 1 if BMI percentile is greater than or equal to 85^{th} percentile and less than 95^{th} percentile, and 0 if not; and the obesity indicator is set to 1 if BMI percentile is greater than or equal to 95^{th} percentile, and 0 otherwise. These three measures are constructed for all seven waves, while another energy balance measure, body fat %, is only available for three waves, i.e. 1999-2000, 2001-2002, and 2003-2004.

The construction of the lifestyle variables, i.e. sugar-sweetened beverage calories intake, other calories intake, and minutes of physical activity per day, is a little tricky as dietary data released and physical activity questionnaire vary across waves. Started from wave 2003-2004, the survey releases two days of calories intake data for each participant. The first day diet recall is collected in the MEC, and the second day recall is collected via telephone 3-10 days later. Most of the participants have two days of intakes available. Therefore, for the waves released since 2003, we use the average of calories intake if two days of intake data are available, and use one day of intake data if not when constructing calories intake variables. For wave 1999-2000 and 2001-2002, calories intake variables are built based on the one day of intake data released. Considering the potential inconsistency of calories intake variables we use across waves, we also control a variable to indicate whether calories intake variables are generated by using two-day data or not. Because of the concern that a 24-hour diet recall may not reflect one's usual diet behavior, e.g. people may eat more over the weekend than on weekdays, so we generate a variable to show the

proportion of diet recall(s) that happened during the weekend.[36] There were Physical activity (PA) questionnaire changes since 2007. Prior to 2007, participants were asked about specific types of leisure time activities (e.g. basketball, baseball, yoga, etc.), for each of which, they were asked about the intensity of the activity, i.e. vigorous or moderate, the number of times in past 30 days, as well as the minutes on average spent each time. Based on this information, we generate three PA variables for any types of moderate, vigorous, and moderate-vigorous combined activities respectively, and they are defined as minutes spent per day. In 2007 and beyond, participants are not asked about specific types of physical activities, but asked about moderate or vigorous physical activities in general, such as minutes spent on moderate/vigorous activities at work on a typical day; minutes spent on walking or bicycling for transportation purpose on a typical day; and minutes spent on moderate/vigorous recreational activities on a typical day. To be consistent with previous waves, we only consider recreational activities for the waves released after 2006 when constructing the relevant PA variables, assuming that leisure time activities are approximately equivalent to recreational activities.

Other control variables, such as age, gender, race, household income, number of people in the household, and reference person's[37] marital status and educational level, are also obtained from the NHANES. And the three publicly available instrumental variables, i.e. school breakfast availability, having family members receiving Food Stamps, and number of times of eating restaurant per week, are also obtained from the NHANES.

[36] This variable can take three values, 0, 0.5, and 1. Value 0 means no diet recall(s) was(were) on a weekend; 0.5 means one of the two-day diet recalls was on a weekend; and 1 indicates that diet recall(s) was(were both) on a weekend.

[37] Reference person is the one who owns or rents the residence where other household members reside. He/she is not necessarily the parent of the child, but may still play an important role in affecting child's lifestyle behaviors. So we control for the characteristics of the reference person when analyzing children's energy balance.

We will merge the pooled data set constructed from the NHANES to a database of aggregate-level "prospective policy levers" based on the state and county identifiers, which are only available in the restricted data available through Census Research Data Centers (RDC). We are currently in the process of getting access to the RDC data. Once we gain access, we will do the merge and use the policy-lever variables as the instrumental variables, instead of the three publicly available ones mentioned above, in the empirical analysis illustrated in section 3.3.2 below. We will also include a richer set of control variables than those used in the parsimonious regression specification used below. We construct the database of policy levers using data from various sources. Our first source of price data is the Council for Community and Economic Research's Cost of Living Index (C2ER COLI); our second source is the United States Department of Agricultures' Quarterly Food at Home Price Database (QFAHPD). Our primary source for numbers of establishments (used to measure the access to restaurants, food stores, etc.) is the Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW), which provides economic data by industry. Data for other policy levers has been obtained directly from, or reconstructed using, databases implemented in published studies.

3.3.2 Illustration of the Proposed Framework – An Empirical Analysis

We use part of the NHANES data discussed above to illustrate the way of implementing our method in practice. Table 3.1 shows the variables used in this illustrative analysis and the summary statistics correspondingly. The analysis sample is constructed from the first 3 waves from the NHANES, i.e. 1999-2000, 2001-2002, and 2003-2004, and consists of 2,828 children aged from 12-19, with 29% body fat on average. Sample means

of the three lifestyle variables – SSB calories intake, other calories intake and physical activity per day – are 263 cal., 1975 cal., and 59 minutes respectively. Instrumental variables include indicator for whether the school serves breakfast everyday, indicator for whether there were any family members receiving Food Stamps in the past 12 months, and the number of times of eating restaurant food per week. Other controls involved are children's age, gender and race; household income; and children's reference person's marital status and educational level. Table 3.2 shows the 2SRI first stage regression results using the NLS method for each of the three lifestyle regressions modeled via equation (3.3). Wald test statistics are reported to show the joint significance of the three instrumental variables in the estimation of lifestyle regression equations: 18.98 ($p < 0.01$), 23.02 ($p < 0.01$) and 13.42 ($p < 0.01$) for SSB calories intake, other calories intake and physical activity per day respectively, indicating that our instrumental variables are relevant.

After the first stage lifestyle regressions, we calculate the residuals correspondingly and use them as extra controls in the second stage energy balance regression, where the energy balance outcome is body fat %. This second stage regression is modeled by equation (3.4) and estimated by the NLS method. The results are displayed in Table 3.3, column (1). As a comparison, we also reported the results in column (2) based on the nonlinear regression (NR) method that ignores the potential endogeneity of lifestyle variables in the body fat % regression, i.e. without including residuals in the regression. The asymptotic t statistics of the 2SRI second stage estimates are adjusted to account for the fact that the residuals controlled in the regression are the generated regressors calculated using the first stage estimates. Derivations of the correct asymptotic standard errors of 2SRI second-stage coefficient estimates are discussed in Appendix 3B.

Our 2SRI second stage results suggest that body fat % increases as other calories intake increases at 5% significance level, while the corresponding NR estimate suggest the opposite. Besides, physical activity is shown to have significantly positive effect on decreasing body fat % based on the NR estimate while it seems not to have any significant impact on body fat % based on the 2SRI estimate. We are pointing these out to show that the estimates obtained from the two methods, i.e. 2SRI and NR, could be very different, and that ignoring the potential endogeneity of the variables of interest could result in very biased estimates A nice feature of the 2SRI method is that it allows us to test for endogeneity directly in the second stage by conducting a joint Wald test of the null hypothesis that the coefficients of the residuals are all equal to zero. The Wald test statistic, i.e. 10.71 ($p = 0.013$), shows that the residuals are jointly significant at 5% significance level, indicating that the three lifestyle variables may be endogenous.

As we've mentioned before, we won't use this analysis to make any inference about the effects of lifestyle on body fat %. Part of the reason is because our instrumental variables may be correlated with the random error term of the energy balance regression equation, and hence, subject to the violation of the IV validity condition. Another reason could be the weak instrument issue. Although the first stage test statistics have shown that our instruments are jointly significant in predicting lifestyle variables, the results in chapter 5, based on models that appear to fit the data better, are not as convincing with regard to the strength of the IVs – especially in the physical activity lifestyle regression. It is primarily these problems with the IVs (invalidity and weakness) that lead us to view the 2SRI results in Tables 3.2 and 3.3 as merely illustrative. The same can be said for the AIE estimates presented in Table 3.4. The AIE on body fat % in response to an increment in

each lifestyle variable is calculated using 2SRI [see column (1)], and NR [see column (2)] coefficient estimates respectively, where the increments we choose are 50, 500 and 30 for SSB calories intake, other calories intake and minutes of physical activity, respectively. One can choose any increments and use equation (3.6) to calculate the corresponding AIEs. AIEs of exogenous changes in the policy variables, i.e. the three instrumental variables in this case, on body fat % can be calculated in a similar way based on equation (3.7). As you can see from table 3.4, AIEs from column (1) and column (2) differ a lot.

3.4 Summary

In this chapter, we extend the simple version of the model to a general one that accommodates multiple policy and lifestyle variables and derive the AIEs in this general case. Using the part of the data we've been able to obtain thus far, we conduct an empirical analysis to demonstrate the implementation of our causal analytic framework in practice. Specifically, we show the regression results for each stage of 2SRI estimation, the IV relevance test in the first stage, the endogeneity test in the second stage, the ultimate AIE estimates and the correct asymptotic standard errors[38] associated with the second stage coefficient estimates and AIE estimators. As a comparison, we also present the results based on the NR method that ignores the endogeneity problem. The empirical results suggest that 2SRI-based estimates and NR-based estimates can differ from each other substantially, and hence draw attention to the importance of accounting for endogeneity. Due to the data limitation, we are not able to provide meaningful results with regard to the causal impacts of lifestyle on children's body fat %. The usage of potentially better instrumental variables, i.e. the aggregate-level "prospective policy levers", requires using

[38] For details, see Appendix 2B and 3C.

state and county identifiers which are not directly available in public. We are in the process of getting access to them from RDC. Once we gain access and, thereafter, link our policy-lever database we've constructed to the NHANES data we've cleaned, we will replicate the analysis presented above with a much richer set of instrumental variables and controls. We expect that these results will yield substantive results that can be used to inform childhood obesity policy.

Appendix 3A

Estimators for the General Model Based on the NR/LIV Method

The NR parameter estimators can be obtained via the following two stages:

Stage 1 – is the same as the one in the 2SRI method, i.e. estimate each lifestyle equation in (3.3) via the nonlinear least squares (NLS) method, and obtain consistent coefficient estimates of the αs (say $\tilde{\alpha}_{oj}^{NR}$, $\tilde{\alpha}_{Pj}^{NR}$);

Stage 2 – estimate energy balance regression (3.4) without including $\underline{X}_u$ by applying the NLS method, and obtain the corresponding estimates of the βs (say $\tilde{\beta}_{Lr}^{NR}$, $\tilde{\beta}_{or}^{NR}$), which are not consistent as the lifestyle variables are incorrectly assumed to be exogenous in this regression equation.

With these NR coefficient estimates in hand, we can now estimate the policy effects. The NR-estimated AIEs of an exogenous increment in the jth lifestyle variable (δ_j) or the kth policy variable (Δ_k) on the rth energy balance measure (EB_r) are

$$\left(\sum_{i=1}^{n}\frac{1}{n}\widetilde{EB_{ri}(\delta_j)}^{NR}\right)-\overline{EB_r} \tag{3A-1}$$

and

$$\left(\sum_{i=1}^{n}\frac{1}{n}\widetilde{EB_{ri}(\Delta_k)}^{NR}\right)-\overline{EB_r} \tag{3A-2}$$

where $\widetilde{EB_{ri}(\delta_j)}^{NR}=\Lambda\left(\underline{\boldsymbol{\mathcal{L}}\left(L_{ji}+\delta_j\right)}\tilde{\beta}_{Lr}^{NR}+\underline{X}_{oi}\tilde{\beta}_{or}^{NR}\right)$, $\underline{\boldsymbol{\mathcal{L}}\left(L_{ji}+\delta_j\right)}$ is the same as $\underline{L}_i$ with its jth element shifted by δ_j; $\widetilde{EB_{ri}(\Delta_k)}^{NR}=\Lambda\left(\underline{\tilde{L}_i(\Delta_k)}^{NR}\tilde{\beta}_{Lr}^{NR}+\underline{X}_{oi}\tilde{\beta}_{or}^{NR}\right)$,

$\underline{\tilde{L}_i(\Delta_k)}^{NR}=\left[\tilde{L}_{1i}(\Delta_k)^{NR}\quad\tilde{L}_{2i}(\Delta_k)^{NR}\quad\tilde{L}_{3i}(\Delta_k)^{NR}\right]$,

$\tilde{L}_{ji}(\Delta_k)^{NR} = \exp(\underline{X}_{oi}\tilde{\alpha}_{oj}^{NR} + \underline{\boldsymbol{\mathcal{P}}(P_{ki}+\Delta_k)}\tilde{\alpha}_{Pj}^{NR}) + \tilde{X}_{uji}^{NR}$, $\tilde{X}_{uji}^{NR} = L_{ji} - \exp(\underline{X}_{oi}\tilde{\alpha}_{oj}^{NR} + P_{ki}\tilde{\alpha}_{Pj}^{NR})$, $\underline{\boldsymbol{\mathcal{P}}(P_{ki}+\Delta_k)}$ is the same as $\underline{P}$ with its kth element replaced by $P_{ki}+\Delta_k$; $\overline{EB}_r$ denotes the sample average for the rth energy balance measure.

The two-stage protocol used for the LIV method is

Stage 1 – estimate each of the linearized lifestyle regressions, $L_j = \underline{X}_o \alpha_{oj} + \underline{P}\alpha_{pj} + u_j$, by OLS, and obtain estimators, $\tilde{\alpha}_{oj}^{LIV}$ and $\tilde{\alpha}_{Pj}^{LIV}$. Then construct the vector $\underline{\tilde{L}}^{LIV} = \left[\tilde{L}_1^{LIV} \quad \tilde{L}_2^{LIV} \quad \tilde{L}_3^{LIV}\right]$, where $\tilde{L}_j^{LIV} = \underline{X}_o \tilde{\alpha}_{oj}^{LIV} + \underline{P}\tilde{\alpha}_{Pj}^{LIV}$.

Stage 2 – estimate the linearized energy balance regression, $EB_r = \underline{\tilde{L}}^{LIV}\beta_{Lr} + \underline{X}_o\beta_{or} + e_r^{LIV}$, by OLS, and obtain estimates of the βs, $\tilde{\beta}_{Lr}^{LIV}$ and $\tilde{\beta}_{or}^{LIV}$.

The LIV-estimated AIEs analogous to (3A-1) and (3A-2) are $\delta_j\tilde{\beta}_{L_jr}^{LIV}$ and $\Delta_k(\tilde{\alpha}_{P_k1}^{LIV}\tilde{\beta}_{L_1r}^{LIV} + \tilde{\alpha}_{P_k2}^{LIV}\tilde{\beta}_{L_2r}^{LIV} + \tilde{\alpha}_{P_k3}^{LIV}\tilde{\beta}_{L_3r}^{LIV})$ respectively, where $\tilde{\beta}_{L_jr}^{LIV}$ is the LIV coefficient estimate of the jth lifestyle variable in the rth energy balance regression equation, and $\tilde{\alpha}_{P_kj}^{LIV}$ is the LIV coefficient estimate of the kth policy variable in the jth lifestyle regression equation.

Appendix 3B: Asymptotic Standard Errors for the 2SRI Coefficient Estimates in the General Model

In this section, we derive the correct asymptotic standard errors of the 2SRI coefficient estimates, where the corresponding two stages are:

Stage 1 – use the nonlinear least squares (NLS) method to estimate each lifestyle equation below

$$L_j = \exp(\underline{X}_o \alpha_{oj} + \underline{P}\alpha_{pj}) + X_{uj} \tag{3B-1}$$

where j = 1, 2, 3, and obtain consistent estimators, $\tilde{\alpha}_j' = \left[\tilde{\alpha}_{oj}' \quad \tilde{\alpha}_{Pj}'\right]$, then calculate the residual as

$$\tilde{X}_{uj} = L_j - \exp(\underline{X}_o \tilde{\alpha}_{oj} + \underline{P}\tilde{\alpha}_{pj}) \tag{3B-2}$$

and construct the residual vector as $\underline{\tilde{X}}_u = \left[\tilde{X}_{u1} \quad \tilde{X}_{u2} \quad \tilde{X}_{u3}\right]$;

Stage 2 – apply the NLS method to the rth energy balance outcome regression equation below

$$EB_r = \Lambda(\underline{L}\beta_{Lr} + \underline{X}_o\beta_{or} + \underline{\tilde{X}}_u\beta_{ur}) + e_r^{2SRI} \tag{3B-3}$$

where r = 1, 2, 3, 4, to obtain consistent estimates of the energy balance parameters $\tilde{\beta}_r' = \left[\tilde{\beta}_{Lr}' \quad \tilde{\beta}_{or}' \quad \tilde{\beta}_{ur}'\right]$.

Therefore, the first stage objective function is

$$\sum_{i=1}^{n} q_1(\alpha, V_i) = \sum_{i=1}^{n} \{q_{11i} + q_{12i} + q_{13i}\} \tag{3B-4}$$

where $q_{1ji} = -\left(L_{ji} - \exp\left(\underline{X}_{oi}\alpha_{oj} + \underline{P}_i\alpha_{Pj}\right)\right)^2$, $V_i = \left[\underline{L}_i \quad \underline{X}_{oi} \quad \underline{P}_i\right]$, $\underline{L}_i = \left[L_{1i} \quad L_{2i} \quad L_{3i}\right]$ and the second stage object function is

$$\sum_{i=1}^{n} q_{2r}(\tilde{\alpha}, \beta_r, Z_{ri}) = -\sum_{i=1}^{n}\left(EB_{ri} - \Lambda(\underline{L}_i\beta_{Lr} + \underline{X}_{oi}\beta_{or} + \tilde{\underline{X}}_{ui}\beta_{ur})\right)^2 \tag{3B-5}$$

where $\tilde{\alpha}' = \left[\tilde{\alpha}_1' \quad \tilde{\alpha}_2' \quad \tilde{\alpha}_3'\right]$, $\beta_r' = \left[\beta'_{Lr} \quad \beta'_{or} \quad \beta'_{ur}\right]$, and $Z_{ri} = \left[EB_{ri} \quad V_i\right]$

Following Terza (2016a, 2016aA, and 2016B), the asymptotic covariance matrix of the first and second stage parameter estimators, i.e. $\tilde{\alpha}$ and $\tilde{\beta}_r$, are

$$\tilde{D}^r = \begin{bmatrix} \tilde{D}^r_{11} & \tilde{D}^r_{12} \\ \tilde{D}^r_{12}' & \tilde{D}^r_{22} \end{bmatrix} \tag{3B-6}$$

where

$$\tilde{D}^r_{11} = \widetilde{AVAR}(\tilde{\alpha})$$

$$\tilde{D}^{r1}_{12} = -\widetilde{AVAR}(\tilde{\alpha})\tilde{E}\left[\nabla_{\beta_r\alpha} q_{2r}\right]'\tilde{E}\left[\nabla_{\beta_r\beta_r} q_{2r}\right]^{-1}$$

$$\tilde{D}^r_{22} = \tilde{E}\left[\nabla_{\beta_r\beta_r} q_{2r}\right]^{-1}\tilde{E}\left[\nabla_{\beta_r\alpha} q_{2r}\right]\widetilde{AVAR}(\tilde{\alpha})\tilde{E}\left[\nabla_{\beta_r\alpha} q_{2r}\right]'\tilde{E}\left[\nabla_{\beta_r\beta_r} q_{2r}\right]^{-1} + \widetilde{AVAR}^*\left(\tilde{\beta}_r\right)$$

$$\tilde{E}\left[\nabla_{\beta_r\alpha} q_{2r}\right] = \frac{\sum_{i=1}^{n} \nabla_{\beta_r}\Lambda(\underline{L}_i\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \tilde{\underline{X}}_{ui}\beta_{ur})'\nabla_{\alpha}\Lambda(\underline{L}_i\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \tilde{\underline{X}}_{ui}\tilde{\beta}_{ur})}{n}$$

$$\tilde{E}\left[\nabla_{\beta_r\beta_r} q_{2r}\right] = \frac{\sum_{i=1}^{n} \nabla_{\beta_r}\Lambda(\underline{L}_i\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \tilde{\underline{X}}_{ui}\tilde{\beta}_{ur})'\nabla_{\beta_r}\Lambda(\underline{L}_i\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \tilde{\underline{X}}_{ui}\tilde{\beta}_{ur})}{n}$$

$$\nabla_{\beta_r}\Lambda(\underline{L}_i\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur})=\Lambda(\underline{L}_i\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur})\cdot$$

$$\left[1-\Lambda(\underline{L}_i\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur})\right]\left[\underline{L}_i \quad \underline{X}_{oi} \quad \underline{\tilde{X}}_{ui}\right]$$

$$\nabla_{\alpha}\Lambda(\underline{L}_i\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur})=\Lambda(\underline{L}_i\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur})\cdot$$

$$\left[1-\Lambda(\underline{L}_i\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur})\right]\left[\tilde{\psi}_{1i} \quad \tilde{\psi}_{2i} \quad \tilde{\psi}_{3i}\right]$$

$$\tilde{\psi}_{ji}=-\beta_{ujr}\exp(\underline{X}_{oi}\,\tilde{\alpha}_{oj}+\underline{P}_i\tilde{\alpha}_{pj})\left[\underline{X}_{oi} \quad \underline{P}_i\right]$$

j = 1, 2, 3, $\widetilde{AVAR}(\tilde{\alpha})$ and $\widetilde{AVAR}^{*}(\tilde{\beta}_r)$ are the estimated covariance matrices obtained from the first and second stage packaged regression results respectively.

Appendix 3C: Asymptotic Standard Errors of the 2SRI-Based Average Incremental Effects

This section is focusing on deriving the asymptotic standard errors of the estimated average incremental effects (AIEs) on the energy balance outcomes in response to an exogenous increment in lifestyle variable (L_j) or a particular SSB-related policy intervention of interest (P_k). Recall the three lifestyle regression equations and four energy balance equations:

$$E\left(L_j \mid \underline{X}_o, \underline{P}\right) = \exp\left(\underline{X}_o\alpha_{oj} + \underline{P}\alpha_{Pj}\right)$$

Lifestyle Regression Equations (3C-1)

$$E\left[EB_r \mid \underline{L}, \underline{X}_o, \underline{X}_u\right] = \Lambda\left(\underline{L}\beta_{Lr} + \underline{X}_o\beta_{or} + \underline{X}_u\beta_{ur}\right)$$

Energy Balance Equations (3C-2)

where $j = 1, 2, 3$; $r = 1, 2, 3, 4$; $\underline{X}_u = \left[X_{u1} \quad X_{u2} \quad X_{u3}\right]$ and $X_{uj} = L_j - \exp\left(X_o\alpha_{oj} + P\alpha_{pj}\right)$. The corresponding estimated average incremental effect of an increment, δ_j, in L_j on a particular energy balance outcome EB_r is

$$\widetilde{PE}_{\delta_j} = \sum_{i=1}^{n} \frac{\widetilde{pe}_{\delta_j i}\left(\tilde{\alpha}, \tilde{\beta}\right)}{n} \tag{3C-3}$$

where

$$\widetilde{pe}_{\delta_j i}\left(\tilde{\alpha}, \tilde{\beta}\right) = \Lambda\left(\underline{\boldsymbol{\mathcal{L}}\left(L_{ji} + \delta_j\right)}\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right) - \Lambda\left(\underline{L}_i\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)$$

$\underline{\mathcal{L}}\left(L_{ji}+\delta_j\right)$ is the same as $\underline{L}_i$ with its jth element shifted by δ_j, and $\underline{\tilde{X}}_{ui}=\left[\tilde{X}_{u1i} \quad \tilde{X}_{u2i} \quad \tilde{X}_{u3i}\right]$, $\tilde{X}_{uji}=L_{ji}-\exp(\underline{X}_{oi}\tilde{\alpha}_{oj}+\underline{P}_i\tilde{\alpha}_{pj})$. In order to derive the asymptotic properties of $\widetilde{PE}_{\delta_j}$, we cast it as a two-stage optimization estimator (2SOE): the first stage comprises consistent estimation of α and β (e.g. via 2SRI) and the second stage is to obtain $\widetilde{PE}_{\delta_j}$ by optimizing the following objective function w.r.t. PE_{δ_j}

$$\sum_{i=1}^{n} q_{\delta_j i}\left(\tilde{\alpha},\tilde{\beta},PE_{\delta_l},Z_i\right) \tag{3C-4}$$

where

$$q_{\delta_j i}\left(\tilde{\alpha},\tilde{\beta},PE_{\delta_j},Z_i\right)=-\left(\widetilde{pe}_{\delta_j i}\left(\tilde{\alpha},\tilde{\beta}\right)-PE_{\delta_j}\right)^2$$

$Z_i=\left[\underline{L}_i \quad \underline{X}_{oi} \quad P_i\right]$ and $\left[\tilde{\alpha}' \quad \tilde{\beta}'\right]$ is the first-stage estimator of $\left[\alpha' \quad \beta'\right]$. Following Terza (2016a, 2016aA, and 2016B), the asymptotic standard error of $\widetilde{PE}_{\delta_j}$ is estimated as

$$\widetilde{AVAR}(\widetilde{PE}_{\delta_j})=\left(\frac{\sum_{i=1}^{n}\nabla_{[\alpha' \quad \beta']}\widetilde{pe}_{\delta_j i}\left(\tilde{\alpha},\tilde{\beta}\right)}{n}\right)\widetilde{AVAR}\left(\left[\tilde{\alpha}' \quad \tilde{\beta}'\right]\right)\left(\frac{\sum_{i=1}^{n}\nabla_{[\alpha' \quad \beta']}\widetilde{pe}_{\delta_j i}\left(\tilde{\alpha},\tilde{\beta}\right)}{n}\right)'$$

$$+\frac{\sum_{i=1}^{n}\left(\widetilde{pe}_{\delta_j i}\left(\tilde{\alpha},\tilde{\beta}\right)-\widetilde{PE}_{\delta_j}\right)^2}{n} \tag{3C-5}$$

where

$$\nabla_{[\alpha' \quad \beta']}\widetilde{pe}_{\delta_j i}(\tilde{\alpha},\tilde{\beta})=\left[\nabla_\alpha\widetilde{pe}_{\delta_j i} \quad \nabla_\beta\widetilde{pe}_{\delta_j i}\right]$$

$$\nabla_\alpha\widetilde{pe}_{\delta_j i}=\left\{\Lambda'\left(\underline{\mathcal{L}}\left(L_{ji}+\delta_j\right)\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)-\Lambda'\left(\underline{L}_i\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)\right\}\cdot$$

$$\left[\psi_{1i} \quad \psi_{2i} \quad \psi_{3i}\right]$$

$$\psi_{ji} = -\tilde{\beta}_{ujr} \exp(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{P}_i\tilde{\alpha}_{Pj})\left[\underline{X}_{oi} \quad \underline{P}_i\right]$$

$$\nabla_\beta \widetilde{pe}_{\delta_{ji}} = \Lambda'\left(\underline{\mathcal{L}\left(L_{ji}+\delta_j\right)}\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \tilde{\underline{X}}_{ui}\tilde{\beta}_{ur}\right)\left[\underline{\mathcal{L}\left(L_{ji}+\delta_j\right)} \quad \underline{X}_{oi} \quad \tilde{\underline{X}}_{ui}\right]$$
$$-\Lambda'\left(\underline{L}_i\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \tilde{\underline{X}}_{ui}\tilde{\beta}_{ur}\right)\left[\underline{L}_i \quad \underline{X}_{oi} \quad \tilde{\underline{X}}_{ui}\right]$$

and $\widetilde{AVAR}\left(\left[\tilde{\alpha}' \quad \tilde{\beta}'\right]\right)$ is the consistent estimate of the asymptotic covariance matrix of $\left[\tilde{\alpha}' \quad \tilde{\beta}'\right]$.

Similarly, the estimated average incremental effect of an increment in the kth policy variable, say Δ_k, on a particular energy balance outcome, EB_r, is

$$\widetilde{PE}_{\Delta_k} = \sum_{i=1}^{n} \frac{\widetilde{pe}_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta}\right)}{n} \tag{3C-6}$$

where

$$\widetilde{pe}_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta}\right) = \Lambda\left(\underline{\tilde{L}_i(\Delta_k)}\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \tilde{\underline{X}}_{ui}\tilde{\beta}_{ur}\right) - \Lambda\left(\underline{L}_i\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \tilde{\underline{X}}_{ui}\tilde{\beta}_{ur}\right)$$

$$\underline{\tilde{L}_i(\Delta_k)} = \left[\tilde{L}_{1i}(\Delta_k) \quad \tilde{L}_{2i}(\Delta_k) \quad \tilde{L}_{3i}(\Delta_k)\right]$$

$$\tilde{L}_{ji}(\Delta_k) = \exp(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{\boldsymbol{\mathcal{P}}(P_{ki}+\Delta_k)}\tilde{\alpha}_{Pj}) + \tilde{X}_{uji}$$

$\underline{\boldsymbol{\mathcal{P}}(P_{ki}+\Delta_k)}$ is the same as $\underline{P}$ with its kth element replaced by $P_{ki}+\Delta_k$, and $\tilde{\underline{X}}_{ui}$ is defined as in (3C-3). $\widetilde{PE}_{\Delta_k}$ can also be considered as a two-stage optimization estimator (2SOE), which can be obtained by 1) first consistently estimating α and β via 2SRI, and 2) second optimizing the following objective function w.r.t. PE_{Δ_k}

$$\sum_{i=1}^{n} q_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta},PE_{\Delta_k},Z_i\right) \tag{3C-7}$$

where

$$q_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta},PE_{\Delta_k},Z_i\right) = -\left(\widetilde{pe}_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta}\right) - PE_{\Delta_k}\right)^2$$

$Z_i = \left[\underline{L}_i \quad \underline{X}_{oi} \quad P_i\right]$ and $\left[\tilde{\alpha}' \quad \tilde{\beta}'\right]$ is the first-stage estimator of $\left[\alpha' \quad \beta'\right]$. The corresponding asymptotic standard error of $\widetilde{PE}_{\Delta_k}$ is estimated as

$$\widetilde{AVAR}(\widetilde{PE}_{\Delta_k}) = \left(\frac{\sum_{i=1}^{n} \nabla_{[\alpha' \quad \beta']}\widetilde{pe}_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta}\right)}{n}\right)\widetilde{AVAR}\left(\left[\tilde{\alpha} \quad \tilde{\beta}\right]\right)\left(\frac{\sum_{i=1}^{n} \nabla_{[\alpha' \quad \beta']}\widetilde{pe}_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta}\right)}{n}\right)'$$

$$+\frac{\sum_{i=1}^{n}\left(\widetilde{pe}_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta}\right) - \widetilde{PE}_{\Delta_k}\right)^2}{n} \tag{3C-8}$$

where

$$\nabla_{[\alpha' \quad \beta']}\widetilde{pe}_{\Delta_k i}\left(\tilde{\alpha},\tilde{\beta}\right) = \left[\nabla_{\alpha}\widetilde{pe}_{\Delta_k i} \quad \nabla_{\beta}\widetilde{pe}_{\Delta_k i}\right]$$

$$\nabla_{\alpha}\widetilde{pe}_{\Delta_k i} = \Lambda'\left(\underline{\tilde{L}_i(\Delta_k)}\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)\left[\nabla_{\alpha 1}\tilde{\lambda}_{1i} \quad \nabla_{\alpha 2}\tilde{\lambda}_{2i} \quad \nabla_{\alpha 3}\tilde{\lambda}_{3i}\right] -$$
$$\Lambda'\left(\underline{L}_i\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)\left[\nabla_{\alpha 1}\tilde{\Pi}_{1i} \quad \nabla_{\alpha 2}\tilde{\Pi}_{2i} \quad \nabla_{\alpha 3}\tilde{\Pi}_{3i}\right]$$

$$\nabla_{\alpha j}\tilde{\lambda}_{ji} = \nabla_{\alpha j}\left\{\tilde{\beta}_{L_j r}\left(\exp(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{\mathcal{P}(P_{ki} + \Delta_k)}\tilde{\alpha}_{Pj}) + \tilde{X}_{uji}\right) + \tilde{\beta}_{ujr}\tilde{X}_{uji}\right\}$$

$$= \tilde{\beta}_{L_j r}\exp(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{\mathcal{P}(P_{ki} + \Delta_k)}\tilde{\alpha}_{Pj})\left[\underline{X}_{oi} \quad \underline{\mathcal{P}(P_{ki} + \Delta_k)}\right]$$

$$-\left(\tilde{\beta}_{L_j r} + \tilde{\beta}_{ujr}\right)\exp\left(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{P}_i\tilde{\alpha}_{Pj}\right)\left[\underline{X}_{oi} \quad \underline{P}_i\right]$$

$$\nabla_{\alpha j}\tilde{\Pi}_{ji} = \nabla_{\alpha j}\tilde{\beta}_{ujr}\left(L_{ij} - \exp\left(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{P}_i\tilde{\alpha}_{Pj}\right)\right)$$

$$= -\tilde{\beta}_{ujr}\exp\left(\underline{X}_{oi}\tilde{\alpha}_{oj} + \underline{P}_i\tilde{\alpha}_{Pj}\right)\left[\underline{X}_{oi} \quad \underline{P}_i\right]$$

$$\nabla_{\beta}\widetilde{pe}_{\Delta_k i} = \Lambda'\left(\underline{\tilde{L}_i(\Delta_k)}\tilde{\beta}_{Lr} + \underline{X}_{oi}\tilde{\beta}_{or} + \underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)\left[\underline{\tilde{L}_i(\Delta_k)} \quad \underline{X}_{oi} \quad \underline{\tilde{X}}_{ui}\right]$$

$$-\Lambda'\left(\underline{L}_i\tilde{\beta}_{Lr}+\underline{X}_{oi}\tilde{\beta}_{or}+\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}\right)\left[\underline{L}_i \quad \underline{X}_{oi} \quad \underline{\tilde{X}}_{ui}\right]$$

Table 3.1 Variables Used in the Illustrative Empirical Analysis – Sample and Summary Statistics

	Mean	**SD**
Energy Balance Outcome (EB **)**		
Body Fat %	0.29	0.11
Lifestyle Variables ($\underline{L}$ **)**		
Sugar-Sweetened Beverage Calories Intake (cal.)	263.42	242.91
Other Calories Intake (cal.)	1974.93	934.14
Physical Activity Per Day (minutes)	59.16	82.75
Other Variables ($\underline{X}_o$ **)**		
Age (years)	15.12	1.91
Female	0.48	0.5
Non-white	0.73	0.45
Annual Household Income < $15,000	0.18	0.38
Reference Person Education-High School Graduate	0.6	0.49
Reference Person Education-Some College	0.26	0.44
Reference Person Education-College Graduate or Higher	0.15	0.36
Reference Person's Marital Status-Single	0.37	0.48
Instrumental Variables ($\underline{P}$ **)**		
School Serve Breakfast Each Day	0.81	0.4
Family Member(s) Receiving Food Stamps in the Last 12 months	0.19	0.39
Number of Times per Week Eating at Restaurant	2.26	1.97
N	2,828	

Table 3.2 2SRI First Stage Estimates

	(1) SSB Calories	(2) Other Calories	(3) Physical Activity
Age (years)	0.080***	0.017***	0.102***
	(9.121)	(3.692)	(6.577)
Female	-0.332***	-0.257***	-0.535***
	(-9.358)	(-15.041)	(-8.976)
Non-white	0.034	-0.046*	-0.073
	(0.705)	(-2.097)	(-1.086)
Annual Household Income < $15,000	-0.120*	0.008	-0.070
	(-1.985)	(0.267)	(-0.779)
Reference Person Education-Some College	-0.034	0.040	0.218**
	(-0.701)	(1.819)	(2.805)
Reference Person Education-College Graduate or Higher	-0.141	0.044	0.118
	(-1.772)	(1.708)	(1.503)
Reference Person is Single	0.067	-0.018	-0.058
	(1.526)	(-0.883)	(-0.897)
School Serve Breakfast Each Day	-0.090	-0.062**	-0.199*
	(-1.475)	(-2.601)	(-2.559)
Food Stamps Received in the Last 12 months	0.019	0.025	0.254*
	(0.295)	(0.956)	(2.491)
# Times/Week Eating at Restaurant	0.036***	0.016***	0.022
	(3.989)	(3.896)	(1.508)
Constant	4.484***	7.477***	2.790***
	(30.309)	(99.791)	(10.896)
IV Relevance Test			
Wald (χ^2)	18.983	23.016	13.423
P-value	< 0.000	< 0.000	0.004
N	2,828	2,828	2,828

t statistics in parentheses
Reference Person Education-High School Graduate is omitted.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.3 2SRI Second Stage and NR Estimates

	(1) 2SRI	(2) NR[39]
SSB Calories (cal.)	-0.003 (-1.848)	-0.00003 (-0.954)
Other Calories (cal.)	0.001* (1.974)	-0.00006 *** (-6.861)
Physical Activity/Day (minutes)	-0.008 (-1.620)	-0.001*** (-3.782)
Age (years)	0.061* (2.551)	-0.007 (-1.726)
Female	0.668** (3.010)	0.553*** (32.319)
Non-white	0.151 (1.600)	0.041* (2.303)
Annual Household Income < $15,000	-0.114 (-1.143)	0.006 (0.306)
Reference Person Education-Some College	-0.095 (-0.961)	-0.064*** (-3.595)
Reference Person Education-College Graduate or Higher	-0.246 (-1.890)	-0.074** (-3.283)
Reference Person is Single	0.070 (0.859	-0.010 (-0.576)
$\tilde{X}_{u1}$ (SSB Calories)	0.003 (1.835)	-
$\tilde{X}_{u2}$ (Other Calories)	-0.001* (-2.074)	-
$\tilde{X}_{u3}$ (Physical Activity)	0.008 (1.524)	-
Constant	-3.188**	-0.895***

[39] For the details about the NR method in the estimation of energy balance outcomes, see Appendix 3A

	(-3.111)	(-12.852)
Endogeneity Test		
Wald (χ^2)	10.708*	-
P-value	0.013	-
N	2,828	2,828

t statistics in parentheses, adjusted for the 2SRI estimates, i.e. column (1).[40]
Reference Person Education-High School Graduate is omitted.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[40] For detailed derivations of the correct asymptotic standard errors for the 2SRI second stage NLS estimates, see Appendix 3B.

Table 3.4 Average Incremental Effects of Lifestyle Variables on Body Fat % -- 2SRI vs NR-Based Estimates

	(1) 2SRI-Based	(2) NR-Based[41]
SSB Calories Intake	-0.031	-0.0003
$(\Delta = 50)$	(-1.914)	(-0.954)
Other Calories Intake	0.131	-0.006***
$(\Delta = 500)$	(1.846)	(-6.902)
Physical Activity	-0.049	-0.003***
$(\Delta = 30)$	(-1.719)	(-3.800)
N	2,828	2,828

t statistics in parentheses, adjusted for the 2SRI-based AIEs in column (1).[42]
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[41] For detailed derivations of the NR-based AIEs, see Appendix 3A.

[42] For detailed derivations of the asymptotic standard errors for the 2SRI-based AIEs, see Appendix 3C. Derivations of the asymptotic standard errors for the NR-based AIEs are essentially the same – eliminate $\underline{\tilde{X}}_{ui}\tilde{\beta}_{ur}$ terms from the calculation of the individual AIE, i.e. equation (3C-3) and use covariance matrix of coefficient estimates obtained based on the NR method, instead of $\widetilde{AVAR}\left(\left[\tilde{\alpha}' \quad \tilde{\beta}'\right]\right)$ in equation (3C-5) would give the correct asymptotic standard errors.

Chapter 4: More Efficient Estimation —

A Full Information Version of the Simplified Model

In the previous two chapters, we have shown that, given the conditional means of the lifestyle variables and the energy balance variables, we can consistently estimate the relevant parameters by applying the two-stage residual inclusion (2SRI) method, where the regressions in both stages are estimated by the nonlinear least squares (NLS) method. If we know the distributions of the lifestyle variables and the energy balance variables, and incorporate this full information in maximum likelihood estimation (MLE) of both stages of a 2SRI protocol, we can obtain parameter estimates that are not only consistent but also more efficient than their NLS counterparts. In this chapter, we examine the potential efficiency gains based on the simple case introduced in chapter 2, using simulated data.

4.1 The Simplified Model Revisited

We assume lifestyle variable (L), soda calorie consumption, conditional on the price variate (P), to be a Generalized Gamma random variable which has the following probability density function (pdf)

$$f(L \mid X_o, P; \kappa, \mu, \sigma) = \frac{\gamma^{\gamma}}{\sigma L \sqrt{\gamma}\,\Gamma(\gamma)} \exp(Z\sqrt{\gamma} - V) \qquad L \geq 0 \qquad (4.1)$$

where X_o and P are defined as in chapter 2; $\Gamma(\ \ )$ is the gamma function; $\gamma = |\kappa|^{-2}$; $Z = \text{sign}(\kappa)\{\ln(L) - \mu\}/\sigma$; $V = \gamma \exp(|\kappa| Z)$; $\mu = X_o \alpha_o + P\alpha_p$; κ, σ, and the αs are the parameters to be estimated. We also suppose that, conditional on L, the observable

confounder age (X_o) and the unobservable confounder (X_u), the energy balance variable, body fat % (EB), is Beta distributed with the pdf

$$h(EB \mid L, X_o, X_u; \xi, \mu) = \frac{\Gamma(\xi)}{\Gamma(\xi\mu)\Gamma(\xi(1-\mu))} EB^{\xi\mu-1} (1 - EB)^{\xi(1-\mu)-1} \qquad 0 < EB < 1 \tag{4.2}$$

where as in chapter 2; $\mu = E[EB \mid L, X_o, X_u] = \Lambda(L\beta_L + X_o\beta_o + X_u\beta_u)$, $\Lambda(\ \ )$ is the logistic cumulative density function (cdf); and ξ and the βs are the parameters to be estimated. It is easy to show that the corresponding conditional means of L and EB are, respectively,

$$EB(L \mid X_o, P) = \exp(X_o\, \alpha_o + P\alpha_p + C) \tag{4.3}$$

where $C = \ln[\kappa^{2\sigma/\kappa} C^*]$, $C^* = \Gamma\{(1/\kappa^2) + (\sigma/k)\} / \Gamma\{1/\kappa^2\}$, and can be absorbed into the intercept component of α_o ; and

$$E[EB \mid L,\ X_o,\ X_u] = \ \Lambda(L\beta_L + X_o\beta_o + X_u\beta_u)\,. \tag{4.4}$$

Based on the two conditional means in (4.3) and (4.4), which are the minimum information needed for NLS estimation, we have shown that the αs and βs can be consistently estimated via the 2SRI method introduced in chapter 2. Unlike chapter 2, we now consider a full information version of the estimation approach. By full information, we mean the conditional probability density functions of L and EB are known, and are given by (4.1)

and (4.2) respectively. As long as the functional forms in (4.1) and (4.2) are correctly specified,[43] the parameters can be consistently estimated by the following two stages:

Stage 1 — use maximum likelihood estimation (MLE) to estimate the lifestyle regression parameters (αs, σ, κ) based on the conditional pdf in (4.1), and obtain consistent estimators, $\tilde{\alpha}^{MLE\prime} = [\tilde{\alpha}_o^{MLE\prime} \quad \tilde{\alpha}_P^{MLE}]$, $\tilde{\sigma}^{MLE}$ and $\hat{\kappa}^{MLE}$, then calculate the residual as

$$\tilde{X}_u^{MLE} = L - \exp(X_o\tilde{\alpha}_o^{MLE} + P\tilde{\alpha}_P^{MLE} + \tilde{C}^{MLE}) \tag{4.5}$$

where

$$\tilde{C}^{MLE} = \ln[(\tilde{\kappa}^{MLE})^{(2\tilde{\sigma}^{MLE}/\hat{\kappa}^{MLE})}\tilde{C}^{MLE*}]$$

$$\tilde{C}^{MLE*} = \Gamma\{(1/(\tilde{\kappa}^{MLE})^2) + (\tilde{\sigma}^{MLE}/\tilde{\kappa}^{MLE})\}/\Gamma\{1/(\tilde{\kappa}^{MLE})^2\}$$

Stage 2 — obtain consistent estimates of the energy balance parameters, $\tilde{\beta}^{MLE\prime} = [\tilde{\beta}_L^{MLE} \quad \tilde{\beta}_o^{MLE\prime} \quad \tilde{\beta}_u^{MLE}]$ and $\tilde{\xi}^{MLE}$, by applying MLE based on the conditional pdf in (4.2) with X_u replaced by $\tilde{X}_u^{MLE}$.

Although the approach introduced in chapter 2 relaxes dependence on the full distributional assumption, unlike the MLE discussed here, the latter may afford substantial gains in efficiency. Moreover, the likelihood formulations in (4.1) and (4.2) are very parametrically flexible so that misspecification bias is less of a concern.

[43] The conditional pdf we assumed for L, equation (4.1), is quite flexible as many distributions, such as Weibull, Log-normal, Exponential, etc., that are commonly used in modeling nonnegative random variables are special cases of General Gamma distribution. The Beta distribution we assumed for EB, equation (4.2), can produce a unimodal, uniform, or bimodal distribution of points that can be either symmetrical or skewed (Paolino, 2001), which is quite flexible. The above-mentioned flexibilities in specifying the conditional pdf for L and EB should largely reduce the likelihood of misspecification bias when applying MLE.

After obtaining consistent MLE parameter estimators from stages 1 and 2, we can calculate the relevant policy effect estimates by substituting the MLE parameter estimates for the $\tilde{\alpha}$s and $\tilde{\beta}$s in equation (2.5) and (2.6). For simplicity, these estimators, together with the parameter estimators introduced in this chapter, will be referred to as MLE estimators; and their counterparts described in chapter 2 will be referred to as NLS estimators. To examine the potential efficiency gains of the MLE estimators relative to the NLS estimators, we conduct a simulation study below.

4.2 Simulation Study — Examine the Potential Efficiency Gains

We generate 1000 samples of size n = 10,000 using the same sampling design as that used to simulate the analysis sample in chapter 2, and to each of them apply two different estimators: (1) minimum information version of the model — apply the nonlinear least squares (NLS) method to estimate the two-stage regressions of 2SRI discussed in chapter 2; (2) full information version of the model — apply maximum likelihood estimation (MLE) for both stages of 2SRI introduced in this chapter. Using the results from each of these models, we estimate the policy effects based on equation (2.5) – (2.6). The results are displayed in Table 4.1. Column 2 lists the true values to be estimated: parameters (the αs and βs, i.e. pre-specified values during data generating process) and policy effects [$AIE(\Delta_L)$ and $AIE(\Delta_P)$, i.e. average incremental effects of an increment in L or P on EB, where the increment is 1].[44] The corresponding NLS and MLE estimators are listed in

[44] True values for the policy effects and recommended policy changes were calculated based on a super sample of 1 million observations generated using the same sampling design as that used to simulate the 1000 replicates each of sample size n = 10,000. See chapter 2 for details about the corresponding true values calculated based on a super sample of size 3 million.

column 3 and 4 respectively. We can see that all the estimators are quite close to their true values, indicating that estimators based on both minimum and full information models are consistent. As our main interest here is the possible efficiency gains by incorporating the full information of the model in the estimation, we calculate mean squared error (MSE) of the relevant estimators, which are presented in column 5 and 6. The comparison of these two columns shows that MLE estimators are more efficient than their NLS counterparts, i.e. the MSE of the latter are much larger. To make the comparison more straightforward, we calculate the percentage decrease in the MSE of each estimator based on the full information model relative to the one based on the minimum information model; and the results are displayed in the last column. As you can see, the efficiency improvement is huge: the percentage decrease in the mean squared error of most MLE estimators, relative to NLS estimators, is more than 50%.

4.3 Summary

In this chapter, we introduce a full information version of the simple model by assuming known forms for the conditional probability density functions of the lifestyle (i.e. soda calorie intake) and energy balance outcome (i.e. body fat %) variables. The regressions in the two stages of the 2SRI protocol are then estimated via the MLE method, which is expected produce more efficient estimates than the NLS based 2SRI method used in in chapter 2. We conduct a simulation study to examine the potential efficiency gains. We find the MLE-based estimators have smaller mean squared error than their NLS-based counterparts, and the percentage gain in efficiency is found to be more than 50% for all the coefficient and AIE estimators.

Table 4.1 Comparison of Minimum and Full Information Versions of the Model — Examine Efficiency Gains

Parameter	True	Estimate		MSE		% of Efficiency Gains
		Minimum Information	Full Information	Minimum Information	Full Information	
α_o	0.001	0.0010638	0.0010007	1.20e-06	3.68e-08	96.93%
α_P	-2	-2.00126	-2.000156	0.0003673	0.0000105	97.14%
β_L	0.007	0 .0070029	0.0070007	3.39e-09	4.29e-10	87.35%
β_o	0.02	0.0200513	0.0200058	6.78e-07	2.32e-08	96.58%
β_u	0.005	0.0049971	0.0049996	3.74e-09	4.95e-10	86.76%
AIE(Δ_L)	0.0010839	0.0010839	0.0010836	1.13e-10	3.65e-11	67.70%
AIE(Δ_P)	-0.1050318	-0.1050567	-0.1050167	3.32e-06	1.54e-06	53.61%

The value of a particular estimator listed in column 3 and 4 is averaged over the 1000 simulated samples, i.e. $\sum_{n=1}^{1000}\frac{1}{1000}\tilde{q}_{jmn}$ where $\tilde{q}_{jmn}$ denotes the jth estimate based on mth model (m = minimum or full information model) obtained from nth sample. Mean square error (MSE) of a particular estimator is measured as $\sum_{n=1}^{1000}\frac{1}{1000}(\tilde{q}_{jmn}-q_j)^2$ where q_j is the true value of $\tilde{q}_{jmn}$. To make the comparison between the minimum and full information models with regard to MSE more straightforward, we calculate the percentage of efficiency gains, listed in the last column, as $\frac{\left|MSE_{full,j}-MSE_{minimum,j}\right|}{MSE_{minimum,j}}\times 100\%$.

Chapter 5: More Efficient Estimation —

A Full Information Version of the General Model

This chapter discusses the full information version of the general model introduced in chapter 3. The idea is the same as that used in chapter 4: both lifestyle and energy balance regression parameters are consistently estimated via maximum likelihood estimation (MLE), giving that their conditional distributions are correctly specified. We expect the MLE parameter estimators and the policy effect estimators calculated from them to be more efficient than their counterparts based on the minimum information model described in chapter 3. As in chapter 3, we complete our discussion with an empirical analysis to demonstrate implementation of the method introduced in this chapter. The analysis is performed using the same data set as that used in chapter 3, and hence is comparable to the empirical analysis conducted in that chapter. Correct asymptotic standard errors for the relevant estimators are derived and coded in Stata®.

5.1 The General Model Revisited

We assume the distributions for the nonnegative continuous lifestyle variables [L_1 – sugar-sweetened beverage (SSB) calories intake; L_2 – other calories intake; L_3 – minutes of physical activity per day] to be generalized gamma (GG) as

$$g_j(L_j \mid \underline{X}_o, \underline{P}) = gg(L_j; \kappa_j, \mu_j, \sigma_j) \tag{5.1}$$

where j = 1, 2, 3; $\underline{X}_o$ and $\underline{P}$ are defined as in chapter 3; gg( ) denotes the generalized gamma pdf which is

$$gg(Y;\kappa,\mu,\sigma)=\frac{\gamma^{\gamma}}{\sigma Y\sqrt{\gamma}\Gamma(\gamma)}\exp(Z\sqrt{\gamma}-V) \qquad Y\geq 0$$

with $\gamma=|\kappa|^{-2}$, $Z=\text{sign}(\kappa)\{\ln(Y)-\mu\}/\sigma$, $V=\gamma\exp(|\kappa|Z)$; $\mu_j=\underline{X}_o\,\alpha_{oj}+\underline{P}\alpha_{pj}$; and κ_j, σ_j, and the α_js are the parameters to be estimated.

The energy balance outcomes we are interested in are BMI percentile (EB_1), body fat % (EB_2), indicators for overweight (EB_3) and obesity (EB_4), among which the first two are fractional variables that take values between 0 and 1, and the last two are binary variables that take values of 0 or 1. We therefore assume EB_1 and EB_2 to be beta distribution; and EB_3 and EB_4 to be Bernoulli distribution. The corresponding conditional probability density functions are, respectively, as follows

$$h(EB_{r1}\mid \underline{L},\underline{X}_o,\underline{X}_u;\xi_{r1},\mu_{r1})=\frac{\Gamma(\xi_{r1})}{\Gamma(\xi_{r1}\mu_{r1})\Gamma(\xi_{r1}(1-\mu_{r1}))}EB^{\xi_{r1}\mu_{r1}-1}(1-EB)^{\xi_{r1}(1-\mu_{r1})-1}$$

$$0<EB_{r1}<1 \qquad (5.2)$$

where r1 = 1, 2, $\underline{L}$ and $\underline{X}_u$ are defined as in chapter 3, $\mu_{r1}=E[EB_{r1}\mid \underline{L},\underline{X}_o,\underline{X}_u]=\Lambda(\underline{L}\beta_{Lr1}+\underline{X}_o\beta_{or1}+\underline{X}_u\beta_{ur1})$, $\Lambda(\ )$ is the logistic cdf, and ξ_{r1} and the β_{r1}s are the parameters to be estimated; and

$$f(EB_{r2}\mid \underline{L},\underline{X}_o,\underline{X}_u;\beta_{r2})=\mu_{r2}{}^{EB_{r2}}\left[1-\mu_{r2}\right]^{1-EB_{r2}} \qquad EB_{r2}=\ 0 \text{ or } 1 \qquad (5.3)$$

where r2 = 3, 4, $\mu_{r2}=E[EB_{r2}\mid \underline{L},\underline{X}_o,\underline{X}_u]=\Lambda(\underline{L}\beta_{Lr2}+\underline{X}_o\beta_{or2}+\underline{X}_u\beta_{ur2})$, and β_{r2}s are the parameters.

To consistently estimate the relevant parameters in this case, we can apply maximum likelihood estimation (MLE) based on the conditional probability density functions in (5.1) – (5.3) in both stages of the 2SRI method.

5.2 Method Illustration – Continued Empirical Analysis

We repeat our real data analysis conducted in chapter 3 by applying MLE instead of NLS to the estimations of lifestyle and energy balance regressions. The correct asymptotic standard errors are derived and calculated for these MLE parameter estimators.[45] The relevant AIE estimators and their correct asymptotic standard errors[46] are re-calculated using MLE parameter estimators and the corresponding correct asymptotic standard errors. For the same reasons as discussed in chapter 3, we view the analysis conducted here as merely illustrative of the econometric framework and method. We are aiming at using this illustrative analysis to give a more concrete perspective of our method, and possibly to demonstrate the feasibility of our method in extended scenarios wherein there are multiple endogenous variables in a nonlinear context.

Columns (1) through (3) of Table 5.1, show the 2SRI first stage MLE coefficient estimates for SSB calories intake, other calories intake and physical activity respectively, wherein the underlying conditional distributions are assumed to be generalized gamma (GG) with probability density functions as defined in (5.1). The Wald test statistics show that the IVs are relevant in predicting SSB calories, i.e. 36.53 ($p < 0.01$), and other calories

[45] See Appendix 5A for details.

[46] The formula is exactly the same as the one used to calculate the asymptotic standard errors of the corresponding AIEs, see equation (3C-5) in Appendix 3C. Replace $[\tilde{\alpha}' \quad \tilde{\beta}']$ with $\left[\tilde{\alpha}^{MLE}{}' \quad \tilde{\beta}^{MLE}{}'\right]$ would give us the correct asymptotic standard errors of the AIEs discussed in this chapter.

intakes, i.e. 23.25 ($p < 0.01$), but not relevant for physical activity, i.e. 3.19 ($p = 0.363$), which is inconsistent with the result obtained in chapter 3, where Wald statistic is 13.423 ($p = 0.004$), see the bottom of Table 3.2, column (3). Given this point of divergence between the NLS (chapter 3) and GG-MLE results in the 2SRI first stage, we conduct a model fit comparison test, i.e. NLS vs GG MLE. The NLS estimation performed in chapter 3 can be equivalently cast as the pseudo maximum likelihood estimator (PMLE) based on the normal distribution with an exponential conditional mean [as in the systematic component of the generic lifestyle equation (3.3)].[47] Unfortunately, the NLS-PMLE is not nested within the GG-MLE so a conventional likelihood ratio test cannot be implemented. For this reason, we use the likelihood ratio (LR) test devised by Vuong (1989) for fit comparisons of likelihood-based non-nested models (see also Wooldridge, 2010, p505-508; Greene, 2012, p534-536). Vuong's LR test statistic (V-LR) is asymptotically normally distributed. In the present context, large negative values of the V-LR indicate rejection of the null hypothesis that the models fit the data equally well in favor of the GG-MLE. Similarly, large positive values support the relative validity of NLS-PMLE. As can be seen at the bottom of Table 5.1, the large negative value of the V-LR for each of the lifestyle regressions indicates that GG-MLE affords better model fit. From this result, we conclude that the seemingly good first stage results regarding the strength of the IVs (based on the chi-squared joint test statistics displayed in Table 3.2), are likely to be misleading. The preferred GG-MLE results give evidence of IV weakness – in particular, regarding their predictive power for minutes of physical activity.

[47]See Gourieroux, Monfort and Trognon, 1984; and Gourieroux and Monfort, 1989, section 8.4.2

This goodness-of-fit analysis also supports our idea of using GG as the conditional distribution for non-negative continuous variables – it is very flexible in terms of the shape of its probability distribution, and hence, may fit the data better. As a comparison, we also present the first stage results based on the LIV method, i.e. the linear instrumental variables method[48], in columns (4) – (6) of Table 5.1. F statistics suggest weak IVs in the estimation of all three lifestyle regressions – none of them are greater than 10, the rule of thumb for IV relevance test in the linear case. This weak IVs conclusion seems to be consistent with the one based on GG MLE. However, it does not indicate that LIV performs better than NLS-based 2SRI method discussed in chapter 3. It could be a coincidence. We have every reason to believe that linear regression is not a good choice especially when the dependent variable has limited value range and its distribution is highly skewed.

Table 5.2 shows the coefficient estimates for body fat % regressions based on the alternative methods. Column (1) shows the estimates obtained from the 2SRI second stage estimation, which is based on the MLE method with the distribution modeled in equation (5.2), i.e. Beta distribution. For simplicity, let's call them "corrected" Beta estimates as they are 2SRI estimates and thus directly account for the potential endogeneity of the lifestyle variables. Conversely, the estimates obtained from the similar Beta regression that ignores endogeneity are referred to as the "uncorrected" Beta estimates – listed in column (2). Column (3) displays the LIV second stage coefficient estimates that are corrected for potential endogeneity but ignore the inherent nonlinearity of the model. Column (4) gives the OLS estimates that ignore both endogeneity and nonlinearity. The AIEs on body fat % in response to an exogenous increment in each lifestyle variable based on these methods

[48] See chapter 2, section 2.4.2 for details.

are displayed in Table 5.3, column (1) – (4) correspondingly. As you can see from Tables 5.2 and 5.3, coefficient and AIE estimates differ substantially across the four model specifications: models that account for the potential endogeneity (i.e. "corrected" Beta and LIV) suggest no significant impacts or significant effects at low significance level (5%, only for physical activity in the "corrected" Beta column) of lifestyle variables on body fat % while the ones that ignore endogeneity (i.e. "uncorrected" Beta and OLS) show significant results at a very high significance level (0.1%, for all three lifestyle variables). Comparison of the AIEs between "corrected" Beta and LIV methods [Table 5.3, column (1) and (3)] also suggests divergent results due to the ignoring nonlinearity even though both methods account for endogeneity. All these comparisons draw our attention to the importance of choosing the appropriate method when dealing with endogeneity in a nonlinear context as results can differ substantially across various methods. As our instruments are weak, we won't draw any inferences from these results. We will replicate the analysis using better IVs, i.e. the aggregate-level "prospective policy levers", once we get access to the restricted RDC data.

5.3 Summary

In this chapter, we discuss the full information model in the general case and apply MLE in both stages of the 2SRI method based on the same data set as that is used in chapter 3. Ideally, we would like to compare the MLE-based estimators obtained from this chapter to the NLS-based ones obtained from chapter 3 and show that the two sets of estimators are similar but the former ones have smaller standard errors, indicating efficiency gains from the fully parametric version of the 2SRI model relative to the minimally parametric

one. However, due to the data limitation discussed above, we are not able to do so. Our MLE-based 2SRI first stage test statistics indicate that the instrumental variables we use are not strong enough, and hence, neither NLS-based estimators nor MLE-based estimators are consistent. Therefore, the comparison of standard errors is futile in the current analysis. We are hoping that, once we merge the public NHANES data to the aggregate-level policy data based on state and county identifiers, we will get meaningful results that allow such comparisons.

Appendix 5A: Asymptotic Standard Errors for 2SRI Coefficient Estimates in the Fully Parametric General Model

In this section, we derive the correct asymptotic standard errors of the 2SRI coefficient estimates, where the corresponding two stages are:

Stage 1 – use the maximum likelihood estimation (MLE) method to estimate parameters of each lifestyle conditional pdf defined as in (5.1), and obtain consistent estimators, $\tilde{\alpha}_j^{MLE\,\prime} = \left[\tilde{\alpha}_{oj}^{MLE\,\prime} \quad \tilde{\alpha}_{Pj}^{MLE\,\prime}\right]$, $\tilde{\sigma}_j^{MLE}$ and $\tilde{\kappa}_j^{MLE}$ then calculate the residual as

$$\tilde{X}_{uj}^{MLE} = L_j - \exp\left(\underline{X}_o\tilde{\alpha}_{oj}^{MLE} + \underline{P}\tilde{\alpha}_{Pj}^{MLE} + \tilde{C}_j^{MLE}\right) \tag{5A-1}$$

where

$$\tilde{C}_j^{MLE} = \ln[(\tilde{\kappa}_j^{MLE})^{(2\tilde{\sigma}_j^{MLE}/\tilde{\kappa}_j^{MLE})}\tilde{C}_j^{MLE*}]$$

$$\tilde{C}_j^{MLE*} = \Gamma\{(1/(\tilde{\kappa}_j^{MLE})^2) + (\tilde{\sigma}_j^{MLE}/\tilde{\kappa}_j^{MLE})\}/\Gamma\{1/(\tilde{\kappa}_j^{MLE})^2\}$$

and construct the residual vector as $\underline{\tilde{X}}_u^{MLE} = \left[\tilde{X}_{u1}^{MLE} \quad \tilde{X}_{u2}^{MLE} \quad \tilde{X}_{u3}^{MLE}\right]$;

Stage 2 – apply MLE based on the conditional pdf

$$h\left(EB_{r1} \mid \underline{L}, \underline{X}_o, \underline{\tilde{X}}_u^{MLE}; \xi_{r1}, \mu_{r1}\right) = \frac{\Gamma(\xi_{r1})}{\Gamma(\xi_{r1}\mu_{r1})\Gamma(\xi_{r1}(1-\mu_{r1}))} EB^{\xi_{r1}\mu_{r1}-1} \cdot$$

$$(1 - EB)^{\xi_{r1}(1-\mu_{r1})-1} \tag{5A-2}$$

where r1 = 1, 2, $\mu_{r1} = E\left[EB_{r1} \mid \underline{L}, \underline{X}_o, \tilde{X}_u^{MLE}\right] = \Lambda\left(\underline{L}\beta_{Lr1} + \underline{X}_o\beta_{or1} + \tilde{X}_u^{MLE}\beta_{ur1}\right)$ and the conditional pdf

$$f\left(EB_{r2} \mid \underline{L}, \underline{X}_o, \underline{\tilde{X}}_u^{MLE}; \beta_{r2}\right) = \mu_{r2}{}^{EB_{r2}}\left[1-\mu_{r2}\right]^{1-EB_{r2}} \tag{5A-3}$$

where r2 = 3, 4, $\mu_{r2} = E\left[EB_{r2} \mid \underline{L}, \underline{X}_o, \tilde{\underline{X}}_u^{MLE} \right] = \Lambda\left(\underline{L}\beta_{Lr2} + \underline{X}_o\beta_{or2} + \tilde{\underline{X}}_u^{MLE}\beta_{ur2} \right)$ to obtain consistent estimates of the energy balance parameters $\tilde{\beta}_{r1}^{MLE\,\prime} = \left[\tilde{\beta}_{Lr1}^{MLE\,\prime} \quad \tilde{\beta}_{or1}^{MLE\,\prime} \quad \tilde{\beta}_{ur1}^{MLE\,\prime} \right]$ and $\tilde{\xi}_{r1}^{MLE}$, and $\tilde{\beta}_{r2}^{MLE\,\prime} = \left[\tilde{\beta}_{Lr2}^{MLE\,\prime} \quad \tilde{\beta}_{or2}^{MLE\,\prime} \quad \tilde{\beta}_{ur2}^{MLE\,\prime} \right]$ for fractional outcomes (i.e. body fat % and BMI percentile) and binary outcomes (i.e. overweight and obesity) respectively.

Therefore, the first stage objective function is

$$\sum_{i=1}^{n} q_1\left(\alpha, V_i\right) = \sum_{i=1}^{n} \left\{ l_{1i} + l_{2i} + l_{3i} \right\} \tag{5A-4}$$

where l_{ji} represents the log-likelihood function of jth lifestyle variable for individual i, $V_i = \left[\underline{L}_i \quad \underline{X}_{oi} \quad \underline{P}_i \right]$, and the second stage object functions are

$$\sum_{i=1}^{n} q_{r1}\left(\tilde{\alpha}^{MLE}, \beta_{r1}, \xi_{r1}, Z_{r1i}\right) = \sum_{i=1}^{n} \left\{ \ln\Gamma\left(\xi_{r1}\right) - \ln\Gamma\left(\xi_{r1}\mu_{r1i}\right) - \ln\Gamma\left(\xi_{r1}\left(1-\mu_{r1i}\right)\right) + \left(\xi_{r1}\mu_{r1i} - 1\right)\ln\left(EB_{r1i}\right) + \left(\xi_{r1}\left(1-\mu_{r1i}\right) - 1\right)\ln\left(1 - EB_{r1i}\right) \right\} \tag{5A-5}$$

where $\tilde{\alpha}^{MLE\,\prime} = \left[\tilde{\alpha}_1^{MLE\,\prime} \quad \tilde{\alpha}_2^{MLE\,\prime} \quad \tilde{\alpha}_3^{MLE\,\prime} \right]$, $\beta_{r1}' = \left[\beta'_{Lr1} \quad \beta'_{or1} \quad \beta'_{ur1} \right]$, $Z_{r1i} = \left[EB_{r1i} \quad V_i \right]$ $\mu_{r1i} = \Lambda\left(\underline{L}_i\beta_{Lr1} + \underline{X}_{oi}\beta_{or1} + \tilde{\underline{X}}_{ui}^{MLE}\beta_{ur1} \right)$, and

$$\sum_{i=1}^{n} q_{r2}\left(\tilde{\alpha}^{MLE}, \beta_{r2}, Z_{r2i}\right) = \sum_{i=1}^{n} \left\{ EB_{r2i}\ln\left(\mu_{r2i}\right) + \left(1 - EB_{r2i}\right)\ln\left(1-\mu_{r2i}\right) \right\} \tag{5A-6}$$

where $\beta_{r2}' = \left[\beta'_{Lr2} \quad \beta'_{or2} \quad \beta'_{ur2}\right]$, $\mu_{r2i} = \Lambda\left(\underline{L}_i\beta_{Lr2} + \underline{X}_{oi}\beta_{or2} + \tilde{\underline{X}}_{ui}^{MLE}\beta_{ur2}\right)$,

$Z_{r2i} = \left[EB_{r2i} \quad V_i\right]$ for fractional outcomes and binary outcomes respectively.

Following Terza (2016a, 2016aA, and 2016B), equation (9), the asymptotic covariance matrix of the first and second stage parameter estimators, i.e. $\tilde{\alpha}^{MLE}$ and $\tilde{\beta}_{r1}^{MLE}$, for fractional energy balance outcomes is[49]

$$\tilde{D}^{r1} = \begin{bmatrix} \tilde{D}_{11}^{r1} & \tilde{D}_{12}^{r1} \\ \tilde{D}_{12}^{r1\,\prime} & \tilde{D}_{22}^{r1} \end{bmatrix} \tag{5A-7}$$

where

$$\tilde{D}_{11}^{r1} = \widetilde{AVAR}\left(\tilde{\alpha}^{MLE}\right)$$

$$\tilde{D}_{12}^{r1} = \widetilde{AVAR}\left(\tilde{\alpha}^{MLE}\right)\tilde{E}\left[\nabla_{\beta_{r1}} q_{r1}' \nabla_{\alpha} q_{r1}\right]'\widetilde{AVAR}^{*}\left(\tilde{\beta}_{r1}^{MLE}\right)$$

$$\tilde{D}_{22}^{r1} = \widetilde{AVAR}^{*}\left(\tilde{\beta}_{r1}^{MLE}\right)\tilde{E}\left[\nabla_{\beta_{r1}} q_{r1}' \nabla_{\alpha} q_{r1}\right]\widetilde{AVAR}\left(\tilde{\alpha}^{MLE}\right)\tilde{E}\left[\nabla_{\beta_{r1}} q_{r1}' \nabla_{\alpha} q_{r1}\right]' \cdot$$

$$\widetilde{AVAR}^{*}\left(\tilde{\beta}_{r1}^{MLE}\right) + \widetilde{AVAR}^{*}\left(\tilde{\beta}_{r1}^{MLE}\right)$$

$$\tilde{E}\left[\nabla_{\beta_{r1}} q_{r1}' \nabla_{\alpha} q_{r1}\right] = \frac{\sum_{i=1}^{n} \nabla_{\beta_{r1}} \tilde{q}_{r1}' \nabla_{\alpha} \tilde{q}_{r1}}{n}$$

$$\nabla_{\beta_{r1}} q_{r1} = \xi_{r1}\nabla_{\beta_{r1}}\mu_{r1i}\left\{-\psi\left(\xi_{r1}\mu_{r1i}\right) + \psi\left(\xi_{r1}\left(1-\mu_{r1i}\right)\right) + \ln\left(EB_{r1i}\right) - \ln\left(1-EB_{r1i}\right)\right\}$$

$$\nabla_{\alpha} q_{r1} = \xi_{r1}\nabla_{\alpha}\mu_{r1i}\left\{-\psi\left(\xi_{r1}\mu_{r1i}\right) + \psi\left(\xi_{r1}\left(1-\mu_{r1i}\right)\right) + \ln\left(EB_{r1i}\right) - \ln\left(1-EB_{r1i}\right)\right\}$$

$$\nabla_{\beta_{r1}}\mu_{r1i} = \Lambda'\left(\underline{L}_i\beta_{Lr1} + \underline{X}_{oi}\beta_{or1} + \underline{X}_{ui}\beta_{ur1}\right)\left[\underline{L}_i \quad \underline{X}_{oi} \quad \underline{X}_{ui}\right]$$

[49] ξ_{r1} is not the parameter of interest. So the correct asymptotic standard error of the corresponding estimate is not covered in equation (5A-7).

$$\nabla_{\alpha}\mu_{r1i} = \Lambda'\left(\underline{L}_i\beta_{Lr1} + \underline{X}_{oi}\beta_{or1} + \underline{X}_{ui}\beta_{ur1}\right)\left[-\beta_{ur11}\exp\left(X_{oi}\alpha_{o1} + P_i\alpha_{P1}\right)\left[X_{oi} \quad P_i\right]\right.$$

$$\left.-\beta_{ur12}\exp\left(X_{oi}\alpha_{o2} + P_i\alpha_{P2}\right)\left[X_{oi} \quad P_i\right] \quad -\beta_{ur13}\exp\left(X_{oi}\alpha_{o3} + P_i\alpha_{P3}\right)\left[X_{oi} \quad P_i\right]\right]$$

$\psi(\ )$ represents the logarithmic derivative of the gamma function, $\widetilde{AVAR}(\tilde{\alpha}^{MLE})$ and $\widetilde{AVAR}^{*}(\tilde{\beta}_{r1}^{MLE})$ are the estimated covariance matrices obtained from the first and second stage packaged regression results respectively. Similarly, for binary outcomes, the asymptotic covariance matrix of both stage parameter estimators, i.e. $\tilde{\alpha}^{MLE}$ and $\tilde{\beta}_{r2}^{MLE}$, is

$$\tilde{D}^{r2} = \begin{bmatrix} \tilde{D}_{11}^{r2} & \tilde{D}_{12}^{r2} \\ \tilde{D}_{12}^{r2}{}' & \tilde{D}_{22}^{r2} \end{bmatrix} \tag{5A-8}$$

where

$$\tilde{D}_{11}^{r2} = \widetilde{AVAR}\left(\tilde{\alpha}^{MLE}\right)$$

$$\tilde{D}_{12}^{r2} = \widetilde{AVAR}\left(\tilde{\alpha}^{MLE}\right)\tilde{E}\left[\nabla_{\beta_{r2}}q_{r2}'\nabla_{\alpha}q_{r2}\right]'\widetilde{AVAR}^{*}\left(\tilde{\beta}_{r2}^{MLE}\right)$$

$$\tilde{D}_{12}^{r2} = \widetilde{AVAR}^{*}\left(\tilde{\beta}_{r2}^{MLE}\right)\tilde{E}\left[\nabla_{\beta_{r2}}q_{r2}'\nabla_{\alpha}q_{r2}\right]\widetilde{AVAR}\left(\tilde{\alpha}^{MLE}\right)\tilde{E}\left[\nabla_{\beta_{r2}}q_{r2}'\nabla_{\alpha}q_{r2}\right]'\cdot$$

$$\widetilde{AVAR}^{*}\left(\tilde{\beta}_{r2}^{MLE}\right) + \widetilde{AVAR}^{*}\left(\tilde{\beta}_{r2}^{MLE}\right)$$

$$\tilde{E}\left[\nabla_{\beta_{r2}}q_{r2}'\nabla_{\alpha}q_{r2}\right] = \frac{\sum_{i=1}^{n}\nabla_{\beta_{r2}}\tilde{q}_{r2}'\nabla_{\alpha}\tilde{q}_{r2}}{n}$$

$$\nabla_{\beta_{r2}}q_{r2} = EB_{r2i}\frac{\nabla_{\beta_{r2}}\mu_{r2i}}{\mu_{r2i}} - \left(1 - EB_{r2i}\right)\frac{\nabla_{\beta_{r2}}\mu_{r2i}}{1-\mu_{r2i}}$$

$$\nabla_{\alpha}q_{r2} = EB_{r2i}\frac{\nabla_{\alpha}\mu_{r2i}}{\mu_{r2i}} - \left(1 - EB_{r2i}\right)\frac{\nabla_{\alpha}\mu_{r2i}}{1-\mu_{r2i}}$$

$$\nabla_{\beta_{r2}}\mu_{r2i} = \Lambda'\left(\underline{L}_i\beta_{Lr2} + \underline{X}_{oi}\beta_{or2} + \underline{X}_{ui}\beta_{ur2}\right)\left[\underline{L}_i \quad \underline{X}_{oi} \quad \underline{X}_{ui}\right]$$

$$\nabla_{\alpha}\mu_{r2i} = \Lambda'\left(\underline{L}_i\beta_{Lr2} + \underline{X}_{oi}\beta_{or2} + \underline{X}_{ui}\beta_{ur2}\right)\left[-\beta_{ur21}\exp\left(X_{oi}\alpha_{o1} + P_i\alpha_{P1}\right)\left[X_{oi} \quad P_i\right]\right.$$

$$\left.-\beta_{ur22}\exp\left(X_{oi}\alpha_{o2} + P_i\alpha_{P2}\right)\left[X_{oi} \quad P_i\right] \quad -\beta_{ur23}\exp\left(X_{oi}\alpha_{o3} + P_i\alpha_{P3}\right)\left[X_{oi} \quad P_i\right]\right]$$

$\widetilde{AVAR}\left(\tilde{\alpha}^{MLE}\right)$ and $\widetilde{AVAR}^{*}(\tilde{\beta}_{r2}^{MLE})$ are the estimated covariance matrices obtained from the first and second stage packaged regression results respectively.

Table 5.1 2SRI and LIV First Stage Estimates

	(1) SSB Calories GG	(2) Other Calories GG	(3) Physical Activity GG	(4) SSB Calories LIV	(5) Other Calories LIV	(6) Physical Activity LIV
Age (years)	0.077***	0.013**	0.119***	20.660***	29.035**	5.451***
	(5.538)	(3.027)	(6.882)	(8.826)	(3.289)	(6.679)
Female	-0.273***	-0.251***	-0.279***	-83.698***	-494.182***	-27.618***
	(-5.513)	(-15.252)	(-4.401)	(-9.729)	(-14.911)	(-9.332)
Non-white	0.064	-0.025	-0.102	13.564	-71.000	-4.987
	(1.165)	(-1.267)	(-1.238)	(1.211)	(-1.704)	(-1.376)
Annual Household Income < $15,000	-0.282***	-0.006	-0.050	-28.530*	3.094	-2.570
	(-4.142)	(-0.235)	(-0.516)	(-2.115)	(0.057)	(-0.596)
Reference Person Education-Some College	0.085	0.036	0.127	-6.576	74.372	10.484**
	(1.514)	(1.796)	(1.679)	(-0.587)	(1.772)	(2.649)
Reference Person Education-College Graduate or Higher	0.191*	0.058*	-0.050	-37.443*	103.288*	5.999
	(2.231)	(2.293)	(-0.504)	(-2.296)	(2.006)	(1.385)
Reference Person is Single	0.175**	-0.025	-0.154*	15.466	-42.759	-4.319
	(3.043)	(-1.323)	(-2.092)	(1.482)	(-1.104)	(-1.346)
School Serve Breakfast Each Day	-0.272***	-0.063**	-0.096	-24.070	-127.222**	-11.495**
	(-4.294)	(-2.966)	(-1.173)	(-1.712)	(-2.634)	(-2.665)

Food Stamps Received in the Last 12 months	0.170*	0.029	0.111	3.072	51.121	9.341
	(2.291)	(1.164)	(1.269)	(0.215)	(0.999)	(1.958)
# Times/Week Eating at Restaurant	0.043***	0.016***	-0.006	12.432***	33.913***	1.274
	(3.409)	(3.698)	(-0.398)	(4.059)	(3.638)	(1.444)
Constant	5.138***	7.521***	3.213***	-21.509	1821.567***	-3.283
	(22.181)	(106.242)	(11.642)	(-0.583)	(12.536)	(-0.260)
ln(σ)	0.157***	-0.832***	0.455***	-	-	-
	(5.338)	(-60.186)	(18.209)			
κ	4.674***	0.452***	3.831***	-	-	-
	(31.374)	(13.247)	(35.760)			
IV Relevance Test						
Wald/F Statistics	36.534	23.246	3.190	6.033	7.135	3.964
P-value	< 0.000	< 0.000	0.363	< 0.000	< 0.000	< 0.008
Model Fit Test (NLS-PMLE vs GG-MLE)						
V-LR Test Statistics	-120145.571	-1293816.814	-103275.558	-	-	-
N	2,828	2,828	2,828	2,828	2,828	2,828

t statistics in parentheses
Wald statistics are reported in column (1) – (3), while F statistics are reported in column (4) – (6).
Reference Person Education-High School Graduate is omitted.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5.2 2SRI Second Stage "Corrected" Beta, "Uncorrected" Beta, LIV Second Stage, and OLS Estimates

	(1) "Corrected" Beta	(2) "Uncorrected" Beta	(3) LIV	(4) OLS
SSB Calories (cal.)	0.001	-0.00003	0.0002	-0.000005
	(0.711)	(-0.880)	(0.164)	(-0.804)
Other Calories (cal.)	-0.001	-0.00006***	-0.00008	-0.00001***
	(-0.999)	(-6.300)	(-0.191)	(-6.946)
Physical Activity/Day (minutes)	-0.015*	-0.001***	0.00002	-0.00009***
	(-1.980)	(-5.064)	(0.007)	(-3.763)
Age (years)	0.059	-0.012**	-0.004	-0.003**
	(1.755)	(-2.774)	(-0.208)	(-2.929)
Female	0.249	0.582***	0.096	0.115***
	(1.251)	(33.978)	(1.434)	(34.506)
Non-white	-0.071	0.038	0.0003	0.008*
	(-1.071)	(1.934)	(0.008)	(2.063)
Annual Household Income < $15,000	0.055	0.013	0.007	0.001
	(0.757)	(0.556)	(0.204)	(0.270)
Reference Person Education-Some College	0.033	-0.064**	-0.008	-0.013***
	(0.487)	(-3.269)	(-0.534)	(-3.482)

Reference Person Education-College Graduate or Higher	-0.095 (-1.032)	-0.072** (-2.876)	-0.002 (-0.033)	-0.015*** (-3.298)
Reference Person is Single	-0.162* (-2.060)	-0.016 (-0.898)	-0.007 (-0.283)	-0.002 (-0.566)
$\tilde{X}_{u1}$ (SSB Calories)	-0.001 (-0.726)	-	-	-
$\tilde{X}_{u2}$ (Other Calories)	0.0004 (0.883)	-	-	-
$\tilde{X}_{u3}$ (Physical Activity)	0.014 (1.914)	-	-	-
Constant	-0.348 (-0.368)	-0.842*** (-11.979)	0.438 (0.533)	0.307*** (22.050)
ln(ξ)	3.248*** (85.531)	3.237*** (123.581)	-	-
Endogeneity Test				
Wald/F Statistics	5.542	-	0.378	-
P-Value	0.136	-	0.769	-
N	2,828	2,828	2,828	2,828

t statistics in parentheses, adjusted for 2SRI second stage "corrected" Beta estimates, i.e. column (1).[50]
Reference Person Education-High School Graduate is omitted.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[50] For detailed derivations of the correct asymptotic standard errors for the 2SRI second stage Beta estimates, see Appendix 5A.

Table 5.3 Average Incremental Effects of Lifestyle Variables on Body Fat %
-- 2SRI-Based "Corrected" Beta vs "Uncorrected" Beta vs LIV vs OLS

	(1) "Corrected" Beta	(2) "Uncorrected" Beta	(3) LIV	(4) OLS
SSB Calories Intake	0.013	-0.0003	0.008	-0.0003
$(\Delta = 50)$	(0.702)	(-.880)	(0.164)	(-0.804)
Other Calories Intake	-0.048	-0.006***	-0.042	-0.006***
$(\Delta = 500)$	(-1.060)	(-6.343)	(-0.191)	(-6.946)
Physical Activity	-0.081*	-0.003***	0.0005	-0.003***
$(\Delta = 30)$	(-2.244)	(-5.083)	(0.007)	(-3.763)
N	2,828	2,828	2,828	2,828

t statistics in parentheses, adjusted for the 2SRI-based AIEs, i.e. column (1) "Corrected" Beta.[51]
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[51] For formula of calculating the correct asymptotic standard error, see equation (3C-5). Replace $[\tilde{\alpha}' \quad \tilde{\beta}']$ with $\left[\tilde{\alpha}^{MLE}' \quad \tilde{\beta}^{MLE}'\right]$ would give the correct asymptotic standard errors of the 2SRI-based AIEs.

Chapter 6: Summary and Discussion

Communities and States are increasingly targeting the consumption of sugar-sweetened beverages (SSBs), especially soda, in their efforts to curb childhood obesity. However, the empirical evidence currently available is not causally interpretable, and hence, provides little or no useful content for policy makers. In the current study, we suggest a modelling framework that can be used for making causal estimation and inference in the context of childhood obesity. This modeling framework is built upon the 2SRI method suggested by Terza et al. (2008), and allows for the implementation of alternative estimation methods at each stage (i.e. NLS or MLE methods, and henceforth NLS-based or MLE-based 2SRI). The framework also accommodates a variety of likelihood specifications (e.g., GG, Beta, logit, etc.), resulting in potentially more efficient estimates. Based on this modeling framework, we derive the estimators that can be used to 1) evaluate the effectiveness of policy interventions on childhood obesity – the average incremental effect (AIE) estimators; and 2) provide quantitative policy recommendations aimed at specified energy balance goals – the policy recommendation estimators. We aim to use those estimators to better inform childhood obesity policy.

We conduct simulation studies in chapter 2 and chapter 4, respectively, 1) to examine the performance of our methods in the estimation of AIEs and quantitative policy recommendations relative to conventional methods – LIV (that ignores inherent nonlinearity) and NR (that does not take account of potential endogeneity); and 2) to assess the potential efficiency gains from implementing MLE-based vs. NLS-based 2SRI. Our simulation studies show that 1) the 2SRI method outperforms LIV and NR methods – estimators obtained from the 2SRI method are very close to the true values while their LIV

and NR counterparts are subject to substantial bias; and 2) MLE-based 2SRI is more efficient than the NLS-based 2SRI approach – the percentage gain in efficiency from MLE vs. NLS is found to be more than 50% for all the coefficient and AIE estimators.

Using publicly available NHANES data, we conduct an empirical study in chapters 3 and chapter 5 to demonstrate the implementation of the methods introduced. We compare the NLS-based 2SRI estimates to their NR counterparts in chapter 3 and find substantial difference in these estimates. In chapter 5 we compare the MLE-based 2SRI estimates to the corresponding estimates obtained from several other alternative methods, including "uncorrected" Beta regression that accounts for nonlinearity but not endogeneity, the LIV method that accounts for endogeneity but not nonlinearity and OLS regression that ignores both endogeneity and nonlinearity. The estimates diverge substantially across different methods. Such findings suggest the importance of choosing the appropriate method when dealing with endogeneity in a nonlinear context. Unfortunately, due to data limitations, we are not able to draw any inference about the causal impacts of lifestyle choices, sugar-sweetened beverage consumption in particular, on childhood obesity. The instrumental variables used in the current empirical analysis are proven to be weak and probably violate the requisite IV validity condition. Potentially better instrumental variables, i.e. the location-related aggregate-level policy variables, will be obtained in the near future. The acquisition of these variables requires the use state and county identifiers which are only available in the Census Research Data Centers (RDC). We are in the process of getting access to the RDC data. Once we get access, we will replicate the empirical analysis performed in chapter 3 and 5 with a much richer set of instrumental variables and controls.

We expect that these results will yield substantive results that can be used to inform childhood obesity policy.

Reference

Anderson, P.M. and Butcher, K.F. (2006): "Childhood Obesity: Trends and Potential Causes." *The Future of Children*, 16, pp. 19-45.

Basu A., and Manca A. (2012): "Regression estimators for generic health-related quality of life and quality-adjusted life-years." *Medical Decision Making*, 32(1): 56-69.

Block, G. (2004): "Foods Contributing to Energy Intake in the US: Data from NHANES III and NHANES 1999–2000." *Journal of Food Composition and Analysis*, 17, pp. 439–447.

Buis, M.L, Cox, J.C., and Jenkins, S.P. (2012): "betafit," http://maartenbuis.nl/software/betafit.html.

Ebbeling, C., Feldman, H., and Osganian, S. (2006): "Effects of Decreasing Sugar-Sweetened Beverage Consumption on Body Weight in Adolescents: A Randomized, Controlled Pilot Study," *Pediatrics*, 117, pp. 673-680.

Flegal, K., Carroll, M., Kuczmarski, R. and Johnson, C. (1998): "Overweight and Obesity in the United States: Prevalence and Trends, 1960-1994," *International Journal of Obesity*, 22, pp. 39-47.

Fletcher, J.M., Frisvold, D. and Tefft, N. (2010a): "Taxing Soft Drinks and Restricting Access to Vending Machines to Curb Child Obesity," *Health Affairs*, 29, pp. 1059-2066.

Fletcher, J.M., Frisvold, D. and Tefft, N. (2010b): "The Effects of Soft Drink Taxes on Child and Adolescent Consumption and Weight Outcomes," *Journal of Public Economics*, 94, pp. 967-974.

Forshee, R., Storey, M., and Ginevan, M. (2005): "A Risk Analysis Model of the Relationship between Beverage Consumption from School Vending Machines and Risk of Adolescent Overweight," *Risk Analysis*, 25, 1121-1135.

Gourieroux, C., Monfort, A., and Trognon, A. (1984): "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, pp. 681-700.

Gourieroux and Monfort (1989): *Statistics and Econometric Models: Volume I*, Cambridge: Cambridge University Press.

Greene, W.H. (2012): *Econometric analysis*, 7th Ed., Pearson Education, Inc.

Healthaliciousness.com (2013): "A List of Common High Glycemic Index (GI) Foods which Can be Eliminated." Available:

http://www.healthaliciousness.com/blog/List-Common-High-Glycemic-Index-GI-foods-which-can-be-eliminated.php.

James, J., Thomas, P., Cavan, D. et al. (2004): "Preventing Childhood Obesity by Reducing Consumption of Carbonated Drinks: Cluster Randomized Controlled Trial." *British Medical Journal*, 328, pp. 1237-1242.

Levy, D., Friend, K., and Wang, Y. (2011): "A Review of the Literature on Policies Directed at the Youth Consumption of Sugar Sweetened Beverages," *Advances in Nutrition*, 2, pp. 182S-200S.

Lin, B., Smith, T.A., Lee, J., and Hall, K.D. (2011): "Measuring weight outcomes for obesity intervention strategies: the case of sugar-sweetened beverage tax," *Economics and Human Biology*, 9:329-341.

Malik, V., Schulze, M., and Hu, F. (2006): "Intake of Sugar-Sweetened Beverages and Weight Gain: A Systematic Review," *American Journal of Clinical Nutrition*, 84, pp. 274-288.

Manning, W.G., Basu, A., Mullahy, J. (2005): "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data," Journal of Health Economics, 24, 245-488.

Mattes RD (1996): "Dietary Compensation by Humans for Supplemental Energy Provided as Ethanol or Carbohydrate in Fluids," *Physiological Behavior*, 59, pp. 179–187.

Ogden, C.L. and Carroll, M.D. (2010): "Prevalence of Overweight, Obesity, and Extreme Obesity Among Adults: United States, Trends 1960-1962 Through 2007-2008," National Center for Health Statistics report. Available: http://www.cdc.gov/nchs/data/hestat/obesity_adult_07_08/obesity_adult_07_08.htm

Ogden, C.L. et al. (2011): "Smoothed percentage body fat percentiles for U.S. children and adolescents, 1999-2004," National Health Statistics Reports, Number 43. Available: http://www.cdc.gov/nchs/data/nhsr/nhsr043.pdf

Ogden, C.L., Carroll, M.D., Kit, B.K., and Flegal, K.M. (2012): "Prevalence of Obesity and Trends in Body Mass Index Among US Children and Adolescents, 1999-2010," *Journal of the American Medical Association*, 307, pp. 483-490.

Paolino, P. (2001): "Maximum Likelihood Estimation of Models with Beta-distributed Dependent Variables," *Political Analysis*, 9, pp. 325-346.

Powell, L. and Chaloupka, F. (2009): "Food Prices and Obesity: Evidence and Policy Implications for Taxes and Subsidies," *Millbank Quarterly*, 87, 229-257.

Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, Pasquale LR, Ridker PM, Hunter DJ, Willett WC, Rimm EB, Chasman DI, Hu FB, Qi L (2012): "Sugar-sweetened beverages and genetic risk of obesity," N Engl J Med. 2012 Oct 11;367(15):1387-96.

Reuters (2013): "Bloomberg's Ban on Big Sodas is Unconstitutional: Appeals Court." Available: http://www.reuters.com/article/2013/07/30/us-sodaban-lawsuit-idUSBRE96T0UT20130730.

Ross, S. (1997): *Simulation, 2nd Ed.*, San Diego: Academic Press.

Stacy, E.W., Mihram, G.A., 1965. Parameter estimation for a generalized gamma distribution. Technometrics 7, 349–358.

Shah, N. (2013): "Rules for School Vending Machines, Snacks Unveiled," Education Week, Available http://blogs.edweek.org/edweek/rulesforengagement/2013/06/rules_for_school_vending_machines_snacks_unveiled.html, Accessed 6/17/14.

Sichieri, R., Paula Trotte, A., de Souza, R., and Veiga, G. (2009): "School Randomised Trial on Prevention of Excessive Weight Gain by Discouraging Students from Drinking Sodas," *Public Health Nutrition*, 12, 197–202.

Sturm, R. (2002): "The Effects of Obesity, Smoking, and Drinking on Medical Problems and costs," *Health Affairs*, 21, pp. 245-253.

Sturm, R., Powell, L., Chriqui, J., and Chaloupka, F. (2010): "Soda Taxes, Soft Drink Consumption, and Children's Body Mass Index," *Health Affairs*, 29, 1052-1058.

Taber, D.R., Stevens, J., Evenson, K.R., Ward, D.S., Poole, C., Maciejewski, M.L., Murray, D.M., and Brownson, R.C. (2011): "State Policies Targeting Junk Food in Schools: Racial Ethnic Differences in the Effect of Policy Change on Soda Consumption," *American Journal of Public Health*, 101, 1769-1775.

Tadikamalla, P.R. (1979): "Random Sampling from the Generalized Gamma Distribution," Computing, 23,199-203**.**

Terza, J.V. and Wu, J. (2016): "Health Policy Analysis from a Potential Outcomes Perspective: Smoking During Pregnancy and Birth Weight," Unpublished Manuscript, Department of Economics, Indiana University Purdue University Indianapolis.

Terza, J.V. (2016a): "Inference Using Sample Means Of Parametric Nonlinear Data Transformations," *Health Services Research*, forthcoming.

Terza, J.V. (2016aA): "Supplementary Appendix to 'Inference Using Sample Means Of Parametric Nonlinear Data Transformations," *Health Services Research*, forthcoming.

Terza, J.V. (2016B): "Simpler Standard Errors for Two-Stage Optimization Estimators", *Stata Journal*, forthcoming.

Terza, J.V., Courtemanche, C.J., Yang, Y., Gupta, S.K., and Chriqui, J. (2014): "The Causal Impact of Sugar Sweetened Beverage Consumption on Childhood Obesity: Evaluating the Effectiveness of Prospective Policy Interventions," Application for Funding, USDA Agriculture and Food Research Initiative, Funding Opportunity USDA-NIFA-AFRI-004492, Tracking Number: GRANT11683665.

Terza, J., Basu, A. and Rathouz, P. (2008): "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling," *Journal of Health Economics*, 27, pp. 531-543.

Terza, J., Bradford, W.D. and Dismuke, C.E. (2008): "The Use of Linear Instrumental Variables Methods in Health Services Research and Health Economics: A Cautionary Note." *Health Services Research*, 43, 1102-1120.

Vartanian, L., Schwartz, M., and Brownell, K. (2007): "Effects of Soft Drink Consumption on Nutrition and Health: A Systematic Review and Meta-Analysis." *American Journal of Public Health*, 97, pp. 667-675.

Vuong, Q.H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica*, 57, pp. 307-333.

Woodward-Lopez, G., Kao, J., and Ritchie, L. (2011): "To What Extent Have Sweetened Beverages Contributed to the Obesity Epidemic?" *Public Health Nutrition*, 14, pp. 499-509.

Wooldridge, J.M. (2010): *Econometric Analysis of Cross Section and Panel Data*, 2nd Ed. Cambridge, MA: MIT Press.

Curriculum Vitae

YAN YANG

Education

Ph.D. Economics, Indiana University, Indianapolis, October 2016

B.A. International Trade and Economics, Anhui University, China, July 2010

Employment

Quantitative Analyst, Sutter Health – Palo Alto Medical Foundation Research Institute, Oct. 2015 – present

Research Expertise

Applied Econometrics, Health Economics

Working Experience

Graduate Assistant, Indiana University Lilly Family School of Philanthropy, Dec. 2014 – Sep. 2015

Research Assistant, Indiana University-Purdue University Indianapolis, Aug. 2012 – May 2013 & Jan. 2014 – June 2014

Teaching Experience

Math Camp (Graduate), Instructor, Indiana University-Purdue University Indianapolis, July 2013 – Aug. 2013

Introduction to Statistical Theory for Economics and Business (Undergraduate), Instructor, Indiana University-Purdue University Indianapolis, June 2015 – Aug. 2015 & Aug. 2013 – Dec. 2013

Statistical Foundations (Graduate), Teaching Assistant, Indiana University-Purdue University Indianapolis, Aug. 2014 – Dec. 2014

Working Paper

"Sugar-Sweetened Beverage Consumption and Childhood Obesity: Formulating Policy Based on Specified Outcome Targets", with Charles Courtemanche and Joseph V. Terza

Programming Skills

STATA, SAS (Certified Advanced Programmer), R, MINITAB, SPSS, Visual Basic

Scholarships, Honors, and Awards

University Fellowship, Indiana University-Purdue University Indianapolis, 2011

Graduation with Honor, Anhui University, China, 2010

National Scholarship for Endeavor (for top 5%), Anhui University, China, 2009