FACULTY PAPER SERIES

FP 01-02 **October, 2000**

The Theory and Econometrics of Health Information in Cross-Sectional Nutrient Demand Analysis

Jaehong Park
George C. Davis

DEPARTMENT OF AGRICULTURAL ECONOMICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS

FACULTY PAPER SERIES

FP 01-02 **October, 2000**

The Theory and Econometrics of Health Information in Cross-Sectional Nutrient Demand Analysis

Jaehong Park
George C. Davis

DEPARTMENT OF AGRICULTURAL ECONOMICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS

Appreciation is extended to all S-278 regional project participants who provided comments on an earlier version of this paper, especially Jeff LaFrance and to Jug Capps, two anonymous reviewers, and the editor for comments that helped improve the paper. Deborah Reed is thanked for providing the grading key used in measuring health information knowledge and Melissa Weichert is thanked for technical assistance.

The Theory and Econometrics of Health Information
in Cross-Sectional Nutrient Demand Analysis

by

Jaehong Park
George C. Davis
gdavis@tamu.edu

Department of Agricultural Economics
Texas A&M University
College Station, Texas 77843-2124

The Theory and Econometrics of Health Information in Cross-Sectional Nutrient Demand Analysis

by

Jaehong Park*
George C. Davis

October, 2000

* Graduate Research Assistant and Associate Professor Department of Agricultural Economics, Texas A&M University.

The Theory and Econometrics of Health Information in Cross-Sectional Nutrient Demand Analysis

Abstract

Understanding the role of health information in food and nutrient demand has become an important issue over the last decade. Endogeneity and measurement error are two empirical problems that are inherent in this type of analysis. While some type of instrumental variables estimation would appear the obvious solution, this paper provides several theoretical and empirical reasons why this is not the case in cross-sectional analysis. An alternative estimation strategy is pursued, an empirical example given, and the implications discussed.

Keywords: Health Information, Demand, Instrumental Variables

JEL: C5, D1, I12

The Theory and Econometrics of Health Information in Cross-Sectional Nutrient Demand Analysis

The perception of food is changing. No longer is it the case that food merely provides taste and sustenance. Rather, medical researchers are rapidly discovering that foods with certain nutrients can act as preventive medicines and this information is quickly reaching the public domain (e.g., Cover Story. Newsweek, November 30, 1998). As this scientific information becomes public knowledge, the potential impacts on the food and medical industries are expected to be substantial. The argument is simple. As consumers become more knowledgeable about the health benefits of certain foods and nutrients, the demand for those foods and nutrients will increase. As the health benefits from consuming those foods and nutrients are realized, the demand for and cost of health care will decline. The obvious public policy corollary then follows: to decrease the future cost of health care, increase the amount of health information knowledge (Fries, Koop, and Beadle). The pivotal implicit assumption in this argument is that health information knowledge has a significant impact on the demand for certain foods or nutrients.

Over the last decade several authors have demonstrated a significant relationship between health information and the consumption of certain foods or nutrients (e.g., Brown and Schrader; Capps and Schmitz; Carlson and Gould; Chern, Loehmann, and Yen; Kinnucan, et al; Gould and Lin; Jensen, Kesavan, and Johnson; Kim, Nayga, and Capps; Variyam, Blaylock, and Smallwood). Of course the significance of the results depend critically on addressing adequately potential econometric problems and in any study of health information knowledge two fundamental econometric problems are especially problematic.

First, since Grossman's seminal work, it has been recognized in the health economics literature that health information is likely an endogenous stock variable that the consumer can

alter through investment decisions. Second, and also widely recognized in the health economics literature, any measure of health information knowledge is likely to be measured with error. An instrumental variables (IV) type estimator is the standard procedure for handling both endogeneity and measurement error and this is usually done in studies using cross-sectional data. However, recent work in the econometrics literature now brings into question this strategy. Because the potential industry and policy implications associated with increasing health information knowledge could be substantial, it is important to explore and understand the limits of the present theory and econometric procedures in determining the impacts of health information knowledge.

This paper addresses three interrelated issues. First, the paper points out some fundamental but overlooked inconsistencies between the theory and econometrics of health information and nutrient demand analysis using instrumental variables techniques. Second, the paper attempts to overcome these inconsistencies by exploiting some recent work found in the econometrics literature. Three, the paper also considers whether the solution is perhaps worse than the original problem. In the next section a rather general theory of nutrient intake behavior is presented. The following section discusses the empirical implications of the theory for the estimation strategy in the light of cross-sectional data constraints. In the following section, the popular Continuing Survey of Food Intake of Individuals (CSFII) and Diet and Health information Survey (DHKS) data is used to empirically explore the empirical consequences following conventional practices. The paper concludes with some suggestions for improving the inferences from cross-sectional and time-series analyses of the relationship between health information and food and nutrient intake.

Theoretical Background and Framework

Grossman's model provides the theoretical foundation for analyzing the demand for health services and his original model is an intertemporal utility maximization problem. Because the emphasis of the paper is on cross-sectional analysis, the theoretical model is simplified here to be a static one period optimization model. This is a common approach in cross-sectional studies (e.g., Pitt, Rosenzweig, and Hassan or Sickles and Taubman). Integrating the models of Becker, Gawn, et al; Pitt, Rosenzweig, and Hassan; Pollak and Wachter; Rosenzweig and Schultz; and Silberberg (1985) adds several refinements to Grossman's model.

Consider a household that produces two final commodities: health services (**H**) and taste (**S**). Health services are produced using the intermediate outputs called nutrients **N** (e.g., cholesterol, fiber), other market inputs $\mathbf{x_H}$ (e.g., medical services or exercise equipment), time $\mathbf{t_H}$, and human capital associated with health knowledge $\mathbf{k_H}$. Nutrient and taste are by their nature joint products, so the technology constraints must reflect this jointness. Consequently, the joint nutrient/taste production technology depends on food inputs $\mathbf{x_f}$, other inputs used in nutrient and taste production $\mathbf{x_N}$ (e.g., cookbooks, utensils), time in preparation and consumption $\mathbf{t_N}$, and human capital representing health knowledge $\mathbf{k_H}$. Health information knowledge is produced and obtained from two general sources: (i) inputs purposely chosen to increase health knowledge regarding nutrient intake $\mathbf{x_k}$ (e.g. a course in nutrition) and (ii) through other market inputs $\mathbf{x_E}$ that generate health information knowledge as an externality through the consumption of other goods (e.g., commercials during entertainment consumption).

Formally, the optimization problem is

(1) Max U = U(**H**, **S**, **x**, **t**) :Utility function (Grossman; Ladd and Suvannunt; Pollak and Wachter)

Subject to

(2.1) $\mathbf{H} = \mathbf{H}(\mathbf{N}, \mathbf{x_H}, \mathbf{t_H}, \mathbf{k_H}; \boldsymbol{\mu})$:Health services production (Grossman; Pitt, Rosenzweig, and Hassan)

(2.2) $\mathbf{G}(\mathbf{N}, \mathbf{S}, \mathbf{x_f}, \mathbf{x_N}, \mathbf{t_N}, \mathbf{k_H}; \boldsymbol{\mu}) = \mathbf{0}$:Nutrient/Taste production (Gawn, et al, Silberberg 1985)

(2.3) $\mathbf{k_H} = \mathbf{K}(\mathbf{x_k}, \mathbf{x_E}, \mathbf{t_H}; \boldsymbol{\mu})$:Health capital knowledge (Rosenzweig and Schultz)

(2.4) $T = \mathbf{t_E} + \mathbf{t_H} + \mathbf{t_N} + \mathbf{t_O} + t_w$:Time Constraint (Becker)

(2.5) $I + p_w t_w = \mathbf{p_x}' \mathbf{x}$:Budget Constraint (Everyone),

where $\mathbf{x} = (\mathbf{x_E}, \mathbf{x_f}, \mathbf{x_H}, \mathbf{x_k}, \mathbf{x_N}, \mathbf{x_O})$ with the corresponding price vector $\mathbf{p_x}$ and $\mathbf{x_O}$ is a vector of other market goods. In addition, $\mathbf{t} = (\mathbf{t_E}, \mathbf{t_H}, \mathbf{t_N}, \mathbf{t_O}, t_w)$ and $\mathbf{t_O}$ is time allocated to other goods and t_w is time allocated to work. The wage rate is p_w, which is conventionally assumed to measure the opportunity cost of time. For notational ease let $\mathbf{p} = (p_w, \mathbf{p_x})$ be the vector of all prices. The variables I and T represent nonwage income and total available time, respectively, and $\boldsymbol{\mu}$ is a vector of demographic/endowment or environmental variables. The theoretical model allows subsets of the **x** and **t** vectors to provide utility directly and indirectly as inputs in the production of **H**, **S**, **N** and $\mathbf{k_H}$ (Pollak and Wachter). Also, the nutrient/taste production function (2.2) is written in implicit form to take into account the joint nature of production.

Assuming a quasi-concave utility function and convex production sets, the problem satisfies the regularity conditions and the optimal solutions are:

(3.1) $\mathbf{x} = \mathbf{x}(\mathbf{p}, I, T; \boldsymbol{\mu})$

(3.2) $\mathbf{t} = \mathbf{t}(\mathbf{p}, I, T; \boldsymbol{\mu})$.

Because these functions are expressed in terms of market prices and not marginal prices, they are well defined regardless of whether or not the technology is constant returns to scale or nonjoint (Pollak and Wachter). Barnett refers to these as "reduced form" equations.

The implied optimal solutions for **H**, **S**, **N** and $\mathbf{k_H}$ are

(4.1) $\mathbf{H} = \mathbf{H}(\mathbf{p}, I, T; \boldsymbol{\mu})$,

(4.2) $\mathbf{S} = \mathbf{S}(\mathbf{p}, I, T; \boldsymbol{\mu})$,

(4.3) $\mathbf{N} = \mathbf{N}(\mathbf{p}, I, T; \boldsymbol{\mu})$,

(4.4) $\mathbf{k_H} = \mathbf{k_H}(\mathbf{p}, I, T; \boldsymbol{\mu})$,

which can be considered reduced form demand functions for the commodities. The indirect utility function would be of the form $V(\mathbf{p}, I, T; \boldsymbol{\mu})$. Using duality theory, it is easy to establish that the partial derivatives of (3.1) through (4.4) cannot be signed because all the exogenous variables are in the constraints (Silberberg 1990 chap. 7). Thus, any sign on a partial derivative is compatible with the theory and compelling ex post explanations of parameter signs should not be confused with the theoretical implications.

Within this optimization framework, because health information knowledge $\mathbf{k_H}$ is considered endogenously determined, nutrient demand is not explicitly a function of health information knowledge. However, health information knowledge $\mathbf{k_H}$ may be actually exogenous or predetermined at some other point in the optimization process. In this case, $\mathbf{k_H}$ would be a quasi-fixed input in the conditional optimization problem. The conditional and unconditional optimization problems are easily reconciled through duality theory (e.g., Cornes) and the Le Chatelier principle.

Let $\mathbf{k_H}$ be considered a quasi-fixed input in the conditional optimization problem, which amounts to ignoring the health capital knowledge constraint (2.3) and assuming interior solutions. In this case, there are conditional demand functions corresponding to (4.3) of the form

(4.3.1) $\overline{\mathbf{N}} = \overline{\mathbf{N}}(\overline{\mathbf{p}}, \overline{\mathrm{I}}, \overline{\mathrm{T}}; \mathbf{k}_{\mathbf{H}}, \boldsymbol{\mu})$,

with the corresponding conditional indirect utility function $\overline{\mathbf{V}} = \overline{\mathbf{V}}(\overline{\mathbf{p}}, \overline{\mathrm{I}}, \overline{\mathrm{T}}; \mathbf{k}_{\mathbf{H}}, \boldsymbol{\mu})$.[1] The overstrike indicates that the price vector and expenditure and time constraints associated with the conditional demand function are subsets of the arguments of the unconditional demand functions (i.e., $\overline{\mathbf{p}} \subset \mathbf{p}, \overline{\mathrm{I}} \subset \mathrm{I}$, and $\overline{\mathrm{T}} \subset \mathrm{T}$). The complement vector of prices $\overline{\mathbf{p}}^{\mathrm{C}}$ (i.e., in set notation $\overline{\mathbf{p}}^{\mathrm{C}} = \mathbf{p} - \overline{\mathbf{p}}$) would consist of the prices of those goods that are used exclusively in the production of $\mathbf{k_H}$. At the optimal point, (4.3) and (4.3.1) are related as,

$$
\begin{aligned}
(4.3.2) \quad \overline{\mathbf{z}}_{\mathbf{N}} &= \overline{\mathbf{z}}_{\mathbf{N}}(\overline{\mathbf{p}}, \overline{\mathrm{I}}, \overline{\mathrm{T}}; \mathbf{k}_{\mathbf{H}}, \boldsymbol{\mu}) \\
&\equiv \mathbf{z}_{\mathbf{N}}(\overline{\mathbf{p}}, \overline{\mathrm{I}}, \overline{\mathrm{T}}; \mathbf{k}_{\mathbf{H}}(\mathbf{p}, \mathrm{I}, \mathrm{T}; \boldsymbol{\mu}), \boldsymbol{\mu}) \\
&= \mathbf{z}_{\mathbf{N}}(\mathbf{p}, \mathrm{I}, \mathrm{T}; \boldsymbol{\mu}),
\end{aligned}
$$

and the indirect utility function and conditional indirect utility function are related as $\mathrm{V}(\mathbf{p}, \mathrm{I}, \mathrm{T}; \boldsymbol{\mu}) = \overline{\mathrm{V}}(\overline{\mathbf{p}}, \overline{\mathrm{I}}, \overline{\mathrm{T}}; \mathbf{k}_{\mathbf{H}}(\mathbf{p}, \mathrm{I}, \mathrm{T}; \boldsymbol{\mu}))$. While the conditional optimization problem permits health information knowledge $\mathbf{k_H}$ to become an argument of the nutrient demand equations in a theoretically consistent fashion, whether or not health information knowledge $\mathbf{k_H}$ is endogenous or exogenous is an empirical question that has important implications for how the empirical model will be estimated.

If health information knowledge $\mathbf{k_H}$ is actually endogenous or measured with error, then estimating the conditional demand (4.3.1) requires some type of instrumental variables (IV) estimator. However, the IV estimator is unnecessary and only yields consistent parameter estimates that are inefficient relative to ordinary least squares (OLS) if health information knowledge is actually predetermined or properly measured. More importantly, the response of

[1] This formulation would be observationally equivalent to a Basmann varying preferences approach to incorporating health information in the utility maximization problem. Consequently, the Basmann model can be considered a special case of Becker's household production theory.

the conditional demand (i.e., with health information knowledge treated as exogenous) with respect to the variables in $\bar{\mathbf{p}}$ and $\boldsymbol{\mu}$ will be different than the response of the unconditional demand (i.e., with health information knowledge treated endogenously) with respect to these same variables, a point recently made by Variyam, et al. Therefore, it is important for policy analysis to determine the most appropriate estimation strategy.

Econometric Implications of the Theory

There are three observations concerning the interaction of cross-sectional data sets and the above theory that at first may seem benign, but they ultimately make the Hausman pre-testing strategy and the gains from IV estimation suspect. One, in cross-sectional data sets there are often no observations on certain desirable variables, such as the price vector **p**. Two, equations (3.1) through (4.4) are all theoretically functions of the same variables. Three, the correlation between variables in cross-sectional data sets is usually low and often less than .3. The interaction of these three facts are at the heart of the estimation difficulties.

Recall the Hausman test is designed to test endogeneity bias and is based on comparing the IV estimator with the OLS estimator. The instruments for the IV estimator must satisfy two conditions: (i) they should be highly correlated with the endogenous/ mismeasured variable (i.e., the *relevance* condition) and (ii) they should be uncorrelated with the disturbance term (i.e., the *exogeneity* condition). Unless *both* of these properties are satisfied, the asymptotic properties of the IV estimator break down (Phillips) and the finite sample properties of the estimator can differ greatly from their ideal asymptotic properties, even in very large samples (Bound, Jaeger, and Baker). To see how these requirements relate to the three observations made above, section 2 of Hall, Rudebusch, and Wilcox is useful background.

Consider the linear model

(5) $y = X\beta + \varepsilon$

where y is a (T×1) vector of observations on the dependent variable, X is a (T×*n*) matrix of regressors having rank equal to *n*, ε is a (T×1) vector of the error process, which has a zero mean and homoskedastic variance $\sigma^2 I_T$, and β is the (*n*×1) vector of unknown parameters. The IV estimator is given by

(6) $\hat{\beta} = (X' P_z X)^{-1} X' P_z y$

where $P_z = Z(Z'Z)^{-1}Z'$ and *Z* is the (T×*k*) matrix of instruments with $k \geq n$. The regularity conditions required for the instrumental variable to be consistent are

(C.1) $T^{-1} X' Z \xrightarrow{p} M_{XZ}$: a finite constant matrix with rank *n*;

(C.2) $T^{-1} Z' Z \xrightarrow{p} M_{ZZ}$: a finite constant matrix with rank *k*;

(C.3) $T^{-1/2} Z'\varepsilon \xrightarrow{D} N(0, \sigma^2 M_{ZZ})$.

Condition C.1 relates to the relevance condition and implies that at a minimum *n* of the *k* instruments must each have some unique explanatory power for the regressors. Condition C.2 requires the instruments be linearly independent and condition C.3 expresses the exogeneity condition. To see why the relevance condition C.1 is especially important, recall the IV estimator can be obtained by minimizing with respect to β

(7) $Q_T(\beta) = [y - X\beta]' P_Z [y - X\beta]$

and following Bowden and Turkington (1984, p. 36)

(8) $T^{-1} Q_T(\beta) = (\beta - \beta_0)' V_T^{-1} (\beta - \beta_0) + o_p(1)$

where β_0 is the true value of β and $V_T = T(X'P_Z X)^{-1}$. Note equation (8) is identified only if the asymptotic limit of V_T exists and is positive definite, and condition C.1 guarantees this, in part. If

the rank condition expressed in C.1 is not satisfied or "close" to not being satisfied, then β_0 is not identified or close to not being identified and the distribution of the IV estimator breaks down.

This econometric theory relates to the three observations made above as follows. Matching the econometric model notation with the theoretical model notation, $y = \overline{\mathbf{N}}$ and $X = (\overline{\mathbf{p}}, \bar{I}, \overline{T}; \mathbf{k}_H, \boldsymbol{\mu})$, so the instrument matrix *Z* must contain variables other than those in X if the rank condition is going to be satisfied. Clearly there are potentially more exogenous variables in the unconditional demand equations (4), which are $(\mathbf{p}, I, T; \boldsymbol{\mu})$, because $\overline{\mathbf{p}} \subset \mathbf{p}$, $\bar{I} \subset I$, and $\overline{T} \subset T$. However, as stated in the first observation, in cross-sectional data sets there often are no observations on prices ($\overline{\mathbf{p}}$ or $\mathbf{p}$), and no information on time or expenditures allocated to all goods not used in the production of health information knowledge (i.e., $\overline{T}$ or $\bar{I}$), so consequently these variables are often not available as instruments and the X matrix is reduced to $X = (I, \boldsymbol{\mu})$.

Can some of the abundantly available demographic\endowment or environmental variables constituting $\boldsymbol{\mu}$ serve as instruments? As indicated by the second observation, all these variables enter all the equations in the same manner. Consequently, using some subset of these variables as instruments implies the rank condition is not satisfied because the rank of *Z* – a subset of X – would not be greater than the rank of X, so legitimate instruments cannot come from the X matrix — theoretically. The real problem is clear. While the theory is extremely explicit about how the choice variables enter the objective and constraint functions, it is extremely vague about how the demographic\endowment or environmental variables enter the objective and constraint functions, and in fact, this theoretical model is much more explicit than many.[2]

[2] This lack of specificity can lead to a feeling of justification for employing some *ad hoc* a priori procedures for omitting certain elements of the environmental vector $\boldsymbol{\mu}$ in some equations and including certain elements in other equations (see discussion below). However, recognizing the

However, suppose one is willing to ignore these internal inconsistencies and selects a subset of μ as instruments. It is at this point where the third observation of low correlation between variables in cross-sectional data comes into play. From the econometric discussion above it is clear that if the correlation between the instruments and the endogenous/mismeasured variable is low, then the parameter is "nearly unidentified" and the IV estimator will have poor finite sample properties and the standard statistical inferences will be very misleading (e.g., Buse; Bound, Jaeger, and Baker; Hall, Rudebusch, and Wilcox; Nakamura and Nakamura; Nelson and Startz; Staiger and Stock). In particular, the IV estimator will be biased in the same direction as the OLS estimator and the loss of efficiency relative to OLS can be substantial. Of particular concern is the fact that when the true coefficient on the endogenous/mismeasured regressor is zero, the IV estimate can be highly significant. This result makes it important to recognize the potential limitations of IV estimators for interpreting the impact of health information knowledge on nutrient demand. However, if this is not bad enough, the low correlation causes more problems.

Nakamura and Nakamura demonstrate that the power of the Hausman statistic is positively related to the correlation between the instruments and the endogenous/mismeasured variable or the degree of relevance. As the instruments become less relevant, the power of the

---

lack of specificity of the theory regarding which elements of μ should enter which equations and hoping to avoid a rather arbitrary choice of instruments, one may be tempted to conduct a statistical search procedure for instruments by looking for variables in μ that are significant in the health information knowledge equations but not significant in the nutrient demand equations and use them as instruments. But a moment's reflection indicates that this strategy leads to an infinite regress because the appropriate estimator to use for the specification search in the nutrient demand equation (i.e., IV) depends on variables (i.e., the instruments) being sought. Furthermore, Hall, Rudebusch, and Wilcox demonstrate that using a statistical search procedure for highly correlated instruments can actually exaggerate the poor properties of the IV estimator. In addition, as Nakamura and Nakamura discuss, if some variables are chosen as instruments that are actually endogenous then there will exist an endogeneity problem even after instrumentation which may be worse than the original endogeneity problem.

Hausman test decreases, so the likelihood of falsely accepting exogeneity increases (i.e., the probability of a Type II error increases). Furthermore, as Nakamura and Nakamura also demonstrate, the Hausman test is a test for the *existence* of endogeneity/measurement error, it is *not a test of the severity* of endogeneity/measurement error bias. Consequently, the Hausman test may be significant and yet the OLS bias relatively small or the Hausman test can be insignificant and the OLS bias relatively large. In addition, given that there are usually several missing variables from the design matrix as stated, the model is inherently misspecified and, as Rhodes and Westbrook show, in such cases OLS is likely to be superior to an IV estimation technique. For these reasons, the gains from pursuing a Hausman pre-test strategy and IV estimation when the instruments are weak becomes questionable.

Recognizing the possible limitations of both the IV and non-IV approaches in such cases, Nakamura and Nakamura recommend the more pragmatic approach of doing out of sample comparisons of the two estimators and looking for some consensus among parameter estimates across different models. Yet to follow this pragmatic advice still requires finding legitimate instruments that do not lead to theoretical and statistical consistency problems. Fortunately, Lewbel recently proposed a procedure that is designed for such situations.

In the next section, Lewbel's procedure is implemented to obtain instruments that avoid some of the theoretical pitfalls mentioned and a Hausman specification test is conducted. Prior to conducting the Hausman specification test, two other specification tests are conducted. First, the R^2 on the instrumental variables regression is checked to provide an indication of the relevance of the instruments and the power of the Hausman test. Second, because as already mentioned it is likely that several important variables are omitted due to data shortcomings, the Godfrey-Hutton testing procedure for distinguishing specification error/instrument problems

from errors-in-variables/ endogeneity is implemented. Out of sample forecasting tests are then conducted to determine whether the IV or OLS estimates are preferred.

Data and Results

The utilized data came from the 1994-1996 Continuing Survey of Food Intake of Individuals (CSFII) and Diet and Health Information Knowledge Survey (DHKS) conducted by the Human Nutrition Information Service of the USDA. These two data sets are rather well known and have been utilized in several studies on nutrition. The CSFII was a multistage, stratified area probability sample of noninstitutionalized individuals in the U.S. The CSFII data includes detailed information about the individuals' socioeconomic variables and nutrient intake over two nonconsecutive days. The DHKS was designed so it could be linked with the CSFII. Around three weeks after CSFII was conducted, adults 20 years and older who completed the day 1 interview in the CSFII were contacted. The sample was designed such that there was no more than one DHKS respondent per household. The DHKS survey asked questions addressing individual knowledge, awareness, and attitude on diet and health issues. The responses to this follow-up survey constitute the DHKS data.

In an attempt to reduce measurement error, the measure of health information knowledge is constructed in a manner similar to Kenkel. In the DHKS, respondents were asked 17 questions pertaining to the relationship between specific nutrients and certain diseases. Each question gives a disease as a possible answer to a question of the general form: What health problems are related to eating too much (little) nutrient A: Disease B? The respondent answered yes or no. For example, one question was, What health problems are related to eating too much fat: cancer? [3]

[3] The seventeen general health problems asked if they were associated with each nutrient intake were: heart disease problems, arthritis problems, bone problems, breathing problems, cancer

Obviously, for some questions the correct answer would be yes and in other questions it would be no. To gauge the accuracy of their individual health information knowledge, a professor of nutrition also completed the survey, and each individual was then assigned a grade from 0 to100 based on the nutritionist's answer key.

Three nutrients are considered here: Fiber, Cholesterol, and Total Fat. Table 1 gives the variable definitions used in the analysis and their means and standard deviations for 1994. The data for 1994 is used in estimating the model and out of sample forecasts comparisons are made for 1995 and 1996. No pretest specification search was conducted to search for significant variables because of the pretest bias problems mentioned and the desire to have an accurate idea of the nominal size of the overall specification test (see discussion below).

In table 1 Fiber, Cholesterol, and Total Fat intake all have standard deviations at least half the size of the mean, indicating a wide range of intake across the sample. In terms of health information knowledge, the average score is highest for Cholesterol (68.25), followed by Total Fat (49.99), and then Fiber (32.92). The average household size is about three, with the age of the main meal planner being almost 50. The average annual income is $33,070 and the average hours of watching television is about three hours a day. The average body mass index is roughly 27. The remaining variables in table 1 are dichotomous variables so they indicate the percentage of the respondents satisfying the stated condition and are rather self-explanatory.

The general specification for estimation is

(5) $y_1 = y_2\beta + \mathbf{X}\Gamma + \varepsilon = \mathbf{W}\delta + \varepsilon$

(6) $y_2 = \mathbf{Z}\Pi + \nu$.

---

problems, colon problems, tooth problems, blood sugar problems, water retention problems, fatigue problems, high cholesterol problems, high blood pressure problems, hyperactivity problems, kidney disease problems, obesity problems, stroke problems, and other problems.

Equation (5) represents the equation for nutrient demand, where y_1 is the nutrient intake, y_2 is health information knowledge and **X** is the matrix of variables considered exogenous. That is, **X** = (1, household size, age, income, tv, bmi, job, college, female, nonwhite, male/female household head, female household head, smoker, special diet, vegetarian, program, disease, region1, region2, region3, central, suburb, quarter2, quarter3). The matrix **W** = (y_2, **X**) and ε is the residual term. Equation (6) is the instrumental variables equation for health information knowledge and, for IV estimation, the rank of **Z** must be greater than the rank of **X**. As discussed, this condition is not satisfied by the theory and available data (i.e., **Z** = **X**).

Lewbel's solution to the problem of insufficient instruments is to use second and third moments of variables as instruments. Following Lewbel, if x_i is an element of the **X** matrix, then $q_1 = (y_1 - \bar{y}_1)(y_2 - \bar{y}_2)$ and $q_i = (x_i - \bar{x}_i)(y_2 - \bar{y}_2)$ are all legitimate instruments, in addition to the x_i variables, and the IV estimator is consistent. In the present context, all continuous variables in the **X** matrix are used to form instruments of this type. This gives one instrument of the q_1 form and five variables of the q_i form for each nutrient equation (i.e., i = 2,3,…,6). So if **Q** = (q_1, q_2,…,q_6), then the instrument matrix is **Z** = (**X**,**Q**) and theoretically satisfies the identification conditions.

Though a 3SLS or system generalized method of moments (GMM) estimator could be used in estimating the equations for Fiber, Cholesterol, and Total Fat, a single equation GMM estimator is implemented here for two reasons. First, the efficiency gains in moving from a single equation estimator to a systems estimator increases as the exogenous variables (instruments) across equations become less correlated. Though there are some instruments that differ across equations, a large number of the instruments are common across equations, so the efficiency gains would be attenuated. Second, and more importantly, the equations are likely to be misspecified to different degrees and a systems estimator will propagate these

misspecifications throughout the entire system. For these reasons, the single equation GMM estimator is implemented in the IV estimation because it automatically accounts for heteroskedasticity by implementing White's heteroskedasticity consistent covariance estimator. In the equations estimated by OLS, heteroskedasticity is also accounted for by implementing White's heteroskedasticity consistent covariance estimator.[4]

The specification tests results are given in table 2. The second column shows the R^2 from regressing the health information knowledge variable on the appropriate instruments. None of the auxiliary R^2 values are above .06, indicating that the instruments are weak and the relevance condition is problematic. Furthermore, these values indicate that the predicted health information knowledge values generated by these regressions will unlikely capture much of the actual variation in the health information knowledge variable. Most importantly, this indicates that the IV estimator will likely have poor sample properties and the power of the Hausman test will be low.

The third column gives the Godfrey/Hutton *J* statistic. This statistic is defined to be $J = N\Re^2$, where N is the sample size and $\Re^2$ is the coefficient of determination from a regression of the IV residual vector $e = y_1 - \mathbf{W}\tilde{\delta}$ on $\mathbf{Z}$, where $\tilde{\delta}$ is the IV estimate of δ. This test also can be considered a Lagrange multiplier test of overidentification (Hausman p. 433) and is the first test in a two step testing procedure. If *J* is large, then the specification of the model and/or the instruments are questionable and the results from the IV estimation are of little value. In this case, the specification and/or instruments need to be reconsidered before conducting the Hausman test. If *J* is small then the next step is to conduct a Hausman test. The *J* statistic is distributed as a Chi-squared distribution with the degrees of freedom equal to the number of

[4] Of course, OLS with White's covariance matrix is equivalent to using a GMM estimator where the variables in the equation serve as their own instruments.

elements in **Q** less the number of endogenous variables in the equation, so in this case there are five degrees of freedom. With five degrees of freedom, the critical value for the *J* statistic is 9.23 at the .10 confidence level and 11.07 at the .05 confidence level. The null hypothesis of no specification/instrument problem is only rejected for the Cholesterol equation, so only the Cholesterol specification seems suspect. However, for comparative purposes the Hausman test also will be conducted for the Cholesterol equation.

The low auxiliary R^2 statistics imply the power of the Hausman test is likely low, so the nominal size of the Hausman test should be increased (Lehmann). In addition, the nested nature of the Godfrey/Hutton testing approach implies that the overall significance level is $\alpha = 1 - (1 - \alpha_J)(1 - \alpha_H)$, where α_J and α_H are the nominal significance levels of the *J* test and Hausman test, respectively.[5] So, for example, if $\alpha_J = .05$ and, following Lehmann's advice letting $\alpha_H = .20$, then the probability of pursuing the wrong estimation strategy is .24 (i.e., making a Type I error). This Type I error would be even higher if one first pursued a specification search for significant variables before implementing the Godfrey-Hutton procedure. The overall conclusion reached from table 2 is that there are endogeneity/measurement error problems in the Fiber and Total Fat equations, but the Cholesterol specification is suspect. Without following the Godfrey/Hutton testing strategy and just conducting the Hausman test, the problematic Cholesterol specification would not have been detected. However, even for the Total Fat and Fiber equations, the Hausman test results only indicate that an endogeneity/ measurement error problem exists. The results do not indicate the *degree* of the OLS bias. For these reasons, the pragmatic advice of Nakamura and Nakamura is followed and both the OLS and IV results are reported.

---

[5] This formula is exact if the tests are independent and Godfrey and Hutton show this is indeed the case.

Table 3 shows the OLS and IV estimates for Fiber, Cholesterol, and Total Fat. Before considering the health information knowledge parameters in some detail, a few general observations can be made. There are a total of 78 OLS parameter estimates and therefore 78 IV parameter estimates in table 3. Of these 78 parameter estimates, 30 of the OLS parameters are significant and 25 of the IV parameters are statistically significant at the 10% level, and 22 of these overlap and have the same sign on the parameter. For those that do not overlap, there are 7 parameter estimates that are significant under OLS but insignificant under IV and 4 parameters that are significant under IV but not significant under OLS. In terms of the signs across estimation methods, of the 78 parameters estimated both ways, 11 differ in sign between OLS and IV. With one exception (nutrient disease knowledge for Fiber), these differences are all associated with insignificant parameter estimates. Furthermore, note, as is common in cross-sectional studies, all of the R^2s are low and with the exception of the OLS model for Total Fat, all R^2s are less .1.

Focusing on the health information knowledge parameter point estimates, they show large discrepancies across estimators, in terms of magnitude and significance. For the Fiber model, the OLS parameter estimate for nutrient/disease knowledge is a positive .05 and significant at the 1% level, while the IV estimate is a negative -.25 and not significant at the 10% level. For the Cholesterol model, a similar result is found where the OLS parameter estimate for nutrient/disease knowledge is a negative -.65 and significant at the 1% level, while the IV estimate is also negative but is about five times as large (-3.54) and is significant at the 10% level. For the Total Fat model, the OLS parameter estimate for nutrient/disease knowledge is -.07 and not significant at the 10% level, but the IV estimate is about 25 times as large (-1.83) and is significant at the 10% level. With the exception of the IV estimate for Fiber, the signs on the parameters concur with intuition: Fiber intake increases with an increase in health information

knowledge and Cholesterol and Total Fat decrease with an increase in health information knowledge.

Perhaps more informative than the point estimates on the parameters are the point and interval estimates on the elasticities with respect to the health information knowledge variables.[6] For Fiber, the OLS elasticity with respect to health information knowledge is .11, with a 95% confidence interval of [.08, .14], whereas the IV elasticity is -.54 with a 95% confidence interval of [-1.12, .04]. For Cholesterol, the OLS elasticity with respect to health information knowledge is -.16 with a 95% confidence interval of [-.24, -.08] and the IV elasticity is -.88 with a 95% confidence interval of [-1.62, -.09]. For Total Fat, the OLS elasticity with respect to health information knowledge is -.05 with a 95% confidence interval of [-.10, -.00] and the IV elasticity is –1.25 with a 95% confidence interval of [-2.44, -.06]. Clearly, and not surprisingly, the IV intervals are wider. For Fiber, there is no overlap in the IV and OLS 95% confidence intervals, but there is some overlap in the IV and OLS 95% confidence intervals for Cholesterol and Total Fat.

If a decision must be made between the two estimators or models, sample forecasts tests can be utilized. There are 1889 and 1858 observations available for 1995 and 1996, respectively, that can be used for out of sample comparisons and table 4 gives the comparisons. The R^2 reported in table 4 is the square of the coefficient of correlation between the actual and fitted values. Overall, the OLS models tend to perform better than the IV models, especially for the Total Fat model. The distance metrics also indicate that OLS is preferred to IV. The root mean square error, mean square percentage error, and mean absolute deviation comparisons all indicate that OLS outperforms IV.

[6] The elasticities are evaluated at the means of the data.

To test whether the OLS forecast are statistically preferred to the IV forecast, a forecast encompassing test was conducted (Harvey, Leybourne, and Newbold). Let e_{OLS} and e_{IV} be the residuals associated with the OLS estimator and the IV estimator, respectively. The forecast encompassing test involves testing the null hypothesis $\lambda = 0$ versus the alternative $\lambda > 1$ in the model $e_{OLS} = \lambda(e_{OLS} - e_{IV}) + \xi$, where ξ is the error of the combined forecast. If the null is not rejected then the OLS forecast encompasses the IV forecast and is preferred to the IV forecast. As Harvey, Leybourne, and Newbold demonstrate, once the sample size exceeds about 250, this hypothesis can be tested as powerfully with the nonparametric Spearman rank correlation test as with any other test. Given both samples have over 1500 observations, this condition is easily satisfied so the forecast encompassing test reported in table 4 is the Spearman rank correlation test. The null hypothesis that the OLS forecast encompasses the IV forecast is not rejected at the 5% level for any of the models and thus the OLS models are statistically superior to the IV models out of sample. Given the weakness of the instruments and resulting weakness of the Hausman pre-testing strategy, if an estimator had to be chosen the results tend to favor OLS estimation over IV estimation, despite the results to the contrary from the Hausman tests in table 2.

A simple policy example can be used to demonstrate the importance of these results. Suppose in a preventive attempt to reduce health costs, the government follows the advice of Fries, Koop, and Beadle and attempts to decrease the cost of coronary heart disease associated with cholesterol intake by increasing health information promotion. Coupling the elasticity estimates for Cholesterol with some estimates and assumptions found in Gray, Malla, and Stephen, a ten percent increase in health information knowledge would decrease coronary heart disease related costs by 3.2 percent based on the OLS point elasticity but by 17.6 based on the IV point elasticity. The corresponding ranges based on the 95% confidence intervals would be to

decrease coronary heart disease related costs between 1.6 and 4.8 percent based on the OLS interval, but between 1.8 and 32.4 percent based on the IV interval.[7] While all the empirical evidence points to the superiority of the OLS estimator, ignoring the internal consistency problems and the relevance condition would lead to pursuing an IV estimation procedure. The policy implications would be that if the policy target was to decrease coronary heart disease cost by some fixed percent by funding health education programs, it would seem likely such programs would be severely *underfunded* or the actual percentage reduction would be severely *below* the target if based on unquestioned IV estimates.

Summary and Conclusions

The purpose of this paper has been to determine the econometric implications that are forthcoming from the economic theory of health information and nutrient demand in a cross-sectional data setting. The rather general household production theory presented here implies that theoretically there are no instrumental variables available in most cross-sectional data sets to correct for endogeneity/measurement error problems and specification searches for instrument candidates are likely to lead to spurious results. To overcome this problem a procedure recently advocated by Lewbel for such situations was implemented. However, the Lewbel instruments proved to be rather weak and in such circumstances the recent econometric literature questions the usefulness of the standard instrumental variables approach. When instruments are weak, the asymptotic properties of the IV estimator are poor and can be misleading. More importantly, the standard Hausman pre-test strategy becomes questionable because the power of the Hausman test is low when the instruments are weak and the Hausman test really only test for the existence of

[7] Based on a medical literature review, Gray, Malla, and Stephen find an elasticity of about 2 between coronary heart disease and cholesterol. They assume that the elasticity between the

OLS bias, not its degree. Given these caveats, the pragmatic advice of Nakamura and Nakamura was followed and both OLS and IV results reported and out of sample forecast comparisons made. If a choice had to be made between the OLS and IV estimators, the OLS results would be preferred despite the significant Hausman test results.

For cross-sectional analysis, a pragmatic way to proceed is to openly acknowledge these theory and data limitations and report multiple model specifications rather than operating under the false pretense of a single specification being correct. In addition, more research needs to be done on the theory of health information and nutrient demand and the advantages and disadvantages of alternative measures of health information knowledge. Regardless, reporting alternative model results will help give an indication of how robust results are across alternative estimators or specifications. The only work we are aware of that takes this progressive and pragmatic approach is the recent work by Variyam et al.

Finally, though the paper focuses on cross-sectional analysis it has implications for time series analysis as well. Many of the empirical difficulties identified here in a cross-sectional analysis are not a factor in time series analysis. Specifically, in time series analysis lagged variables, such as quantities and prices, often can be used as instruments and low R^2s are not a problem in time series analysis. Consequently, instrumental variable procedures could be useful in such situations. However, most of the time series analyses that have appeared in the agricultural economics literature have not treated health information as endogenous or measured with error. This observation can likely be explained by comparing the theoretical frameworks across time series and cross-sectional studies. In general, the cross-sectional studies tend to implement theoretical frameworks that are much more sophisticated, better connected with the health economics literature, and treat health information as endogenous and/or measured with

---

cases of coronary heart disease and the costs of coronary heart disease is 1.

error. Alternatively, the theoretical frameworks found in the times series studies usually just augment a classical static demand system with a health information variable that is treated as exogenous or a preference shifter and not measured with error. This leads to a somewhat paradoxical observation regarding the different estimation procedures used in time series and cross-sectional studies: where instrumental variables type estimators may be most useful (time series data) they have not been used and where they are likely to be less useful they have been used (cross-sectional data). Though the fundamental empirical problems may be similar across time series and cross-sectional studies, the solutions are not.

References

Barnett, W. "Pollak and Wachter on the Household Production Function Approach: Comment." *Journal of Political Economy*. 85(1977):1073-1086.

Basmann, R.L. " A Theory of Demand with Variable Consumer Preferences." *Econometrica*. 24(January, 1956): 47-58.

Becker, G. S. *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. 2nd edition. New York: Columbia University Press for the National Bureau of Economic Research. 1975.

Bound, J., D.A. Jaeger, and R.M. Baker. "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association*. 90(June, 1995): 443-50.

Bowden, R. J. and D. A. Turkington. *Instrumental Variables*. Cambridge: Cambridge University Press. 1984.

Brown, D.J., and L.F. Shrader. "Cholesterol Information and Shell Egg Consumption." *American Journal of Agricultural Economics*. 72(August, 1990): 548-55.

Buse, A. "The Bias of Instrumental Variables Estimators." *Econometrica*. 60(1992):173-180.

Capps, O., Jr., and J.D. Schmitz. "A Recognition of Health and Nutrition factors in Food Demand Analysis." *Western Journal of Agricultural Economics.* 16(July, 1991):21-35.

Carlson, K.A., and B.W. Gould. " The Role of Health information in Determining Dietary Fat Intake." *Review of Agricultural Economics*. 16(1994):373-86.

Chern, W.S., E.T. Loehman, and S.T. Yen. “Information, Health Risk Beliefs, and the Demand for Fats and Oils.” *Review of Economics and Statistics*. 77(August, 1995):555-64.

Cornes, R. *Duality and Modern Economics*. Cambridge: Cambridge University Press. 1992.

Fries, J.F., C. E. Koop, and C. E. Beadle. “Reducing Health Care Costs by Reducing the Need and Demand for Medical Services.” *The New England Journal of Medicine*. 329(July, 1993): 321-5.

Gawn, G., R. Innes, G. Rausser, and D. Zilberman. “Nutrient Demand and the Allocation of Time: Evidence from Guam.” *Applied Economics*. 25(1993):811-30.

Godfrey, L.G., and J.P. Hutton. “Discriminating Between Errors-In-Variables/Simultaneity and Misspecification in Linear Regression Models.” *Economics Letters*. 44(1994): 359-64.

Gould, B.W. and H.C. Lin. “Nutrition Information and Household Dietary Fat Intake.” *Journal of Agricultural and Resource Economics*. 19(December, 1994):349-65.

Gray, R., S. Malla, and A. Stephen. “Canadian Dietary Fat Substitutions, 1955-93, and Coronary Heart Disease Costs.” *Canadian Journal of Agricultural Economics*. 46(1998):233-246.

Grossman, M. *The Demand for Health*: *A Theoretical and Empirical Investigation*. National Bureau of Economic Research. New York: Columbia University Press. 1972.

Hall, A.R., G.D. Rudebusch, and D.W. Wilcox. “Judging Instrument Relevance In

Instrumental Variables Estimation." *International Economic Review*. 37(May, 1996): 283-298.

Harvey, D.I., S.J. Leybourne, and P. Newbold. "Tests for Forecast Encompassing." *Journal of Business & Economic Statistics*. 16(April, 1998): 283-298.

Hausman, J. A. "Specification and Estimation of Simultaneous Equation Models," in *Handbook of Econometrics*. Vol. 1. Z. Griliches and M. Intriligator, Eds. New York: Elsevier. 1983.

Jensen, H.H., T. Kesavna, and S.R. Johnson. "Measuring the Impact of Health Awareness on Food Demand." *Review of Agricultural Economics*. 14(July, 1992): 299-312.

Kenkel, D. "Consumer Health Information and the Demand for Medical Care." *Review of Economics and Statistics* (November, 1990):587-95.

Kim, S., R. M. Nayga, and O. Capps, Jr. "Consumer Use of Nutritional Labels:Endogenous Informaiton and Sample Selection Effects." Paper presented at the S-278 Regional Project on Demand Analysis. October 12-14, 1997.

Kinnucan, H.W., H. Xiao, C-J.Hisa, and J.D. Jackson. "Effects of Health Information and Generic Advertising on U.S. Meat Demand." *American Journal of Agricultural Economics*. 79(February, 1997):13-23.

Ladd, G.W., and V. Suvannunt. "A Model of Consumer Goods Characteristics." *American Journal of Agricultural Economics*. 58(August, 1976):504-10.

Lehmann, E. L. *Testing Statistical Hypotheses*. 2nd ed. New York: Springer. 1986.

Lewbel, A. "Constructing Instruments for Regressions with Measurement Error

When No Additional Data are Available, With an Application to Patents and R&D." *Econometrica*. 65(September 1997): 1201-14.

Nakamura, A., and M. Nakamura. "Model Specification and Endogeneity." *Journal of Econometrics*. 83(March, 1998):213-37.

Nelson, C. R. and R. Startz. "The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument Is a Poor One." *Journal of Business*. 63(1990a):125-140.

_____. "Some Further results on the Exact Small Sample Properties of the Instrumental Variables Estimator." *Econometrica*. 58(1990b):967-976.

Phillips, P.C.B. "Partially Identified Econometric Models," *Econometric Theory* 5(1989): 181-240.

Pitt, M., M. Rosenzweig, and N. Hassan. "Productivity, Health, and Inequality in the Intrahousehold Distribution of Food in Low Income Countries." *American Economic Review*. 80(December, 1990):1139-56.

Pollak, R.A and M. L. Wachter. "The Relevance of the Household Production Function and its Implications for the Allocation of Time." *Journal of Political Economy*. 83(1975):255-77.

Rhodes, G. F. and M. D. Westbrook. "A Study of Estimator Densities and Performance Under Misspecification." *Journal of Econometrics*. 16(1981):311-337.

Rosenzweig, M. R. and T. P. Shultz. "Schooling, Information, and Nonmarket Productivity: Contraceptive Use and Its Effectiveness." *International Economic Review*. Vol. 30. 2(May, 1989): 457-477.

_____. "Estimating a Household Production Function: Heterogeneity, The Demand for Health Inputs, and Their Effects on Birth Weight Sickles, R. C. and P. Taubman. "Mortality and Morbidity Among Adults and the Elderly." in *Handbook of Population and Family Economics*. Vol. 1A. M. Rosenzweig and O. Stark, Eds. New York: Elsevier: 1997.

Silberberg, E. "Nutrition and the Demand for Tastes," *Journal of Political Economy*. 93(October, 1985):881-900.

_____. *The Structure of Economics*: *A Mathematical Analysis*. New York: McGraw Hill. 1990.

Staiger, D. and J. H. Stock. "Instrumental Variables Regression with Weak Instruments." *Econometrica*. 65(May, 1997): 557-586.

Variyam, J. N., J. Blaylock, and D. Smallwood. "A Probit Latent Variable Model Of Nutrition Information and Dietary Fiber Intake." *American Journal of Agricultural Economics*. 78(August, 1996): 628-39.

_____. "Information Effects of Nutrient Intake Determinants on Cholesterol Consumption." *Journal of Agricultural and Resource Economics*. 23(July, 1998):110-125.

Variyam, J. N., J. Blaylock, D. Smallwood, and P.B. Basiotis. *USDA's Healthy Eating Index and Nutrition Information*. Technical Bulletin No. 1866. U.S. Department of Agriculture. Economic Research Service. Washington, D.C. 1998.

Table 1. Variables, Definitions, and Summary Statistics for 1994*

Variable	Definitions and Units	Mean and Standard Deviation
Fiber	Two day average intake of Fiber in grams	15.48 (7.91)
Cholesterol	Two day average intake of Cholesterol in milligrams	267.32 (183.13)
Total Fat	Two day average intake of Total Fat in grams	73.22 (38.24)
Fiber/disease knowledge	Grade on Fiber/Disease questions relative to nutritionist	32.92 (23.83)
Cholesterol/disease knowledge	Grade on Cholesterol/Disease questions relative to nutritionist	68.25 (26.11)
Total fat/disease knowledge	Grade on Total fat/Disease questions relative to nutritionist	49.99 (19.94)
Household size	Number of members of household	2.63 (1.47)
Age	Age of main-meal planner in years	49.604 (17.31)
Income	Total household income in $1000	33.07 (25.37)
TV	Average hours of TV watching per day	2.77 (2.23)
Body mass index	Ratio of body weight in kilograms to height squared in meters	26.38 (5.29)
Job	1 if employed in part or full time job; zero otherwise	.55 (.49)
College	1 if attending school beyond 12^{th} grade; zero otherwise	.43 (.49)
Female		.51 (.50)

Table 1. Variables, Definitions, and Summary Statistics for 1994 (Cont.)

Variable	Definitions and Units	Mean and Standard Deviation
Non-White	1 if Non-white; zero otherwise	.18 (.39)
Male-Female Household Head	1 if male and female head household; zero otherwise	.69 (.48)
Female Household Head	1 if female head household; zero otherwise	.23 (.42)
Smoker	1 if smoker; zero otherwise	.26 (.44)
Special Diet	1 if on a special diet; zero otherwise	.18 (.38)
Vegetarian	1 if a vegetarian; zero otherwise	.03 (.17)
Program	1 if participate in food assistance program; zero otherwise	.11 (.31)
Disease		.42 (.49)
Region 1	1 if in Northwest ???; zero otherwise	.19 (.39)
Region 2	1 if in Southeast ???; zero otherwise	.27 (.45)
Region 3	1 if in Northeast ???; zero otherwise	.34 (.47)
Central	1 if in Central/metropolitan area; zero otherwise	.33 (.47)
Suburb	1 if in Suburb area; zero otherwise	.41 (.49)
Q1	1 if interviewed in first quarter; zero otherwise	.21 (.41)

Table 1. Variables, Definitions, and Summary Statistics for 1994 (Cont.)

Variable	Definitions and Units	Mean and Standard Deviation
Q2	1 if interviewed in second quarter; zero otherwise	.24 (.43)
Q3	1 if interviewed in third quarter; zero otherwise	.28 (.45)

Sample Size is 1778 in 1994 and standard deviations are in parentheses.

Table 2. Summary Statistics*

Variable	1994	1995	1996
Observations	1778	1934	1896
Fiber	15.43	15.76	16.38
	(7.88)	(8.88)	(9.03)
Cholesterol	267.07	269.03	272.46
	(182.43)	(187.96)	(184.77)
Total Fat	72.84	71.92	74.48
	(38.21)	(38.61)	(39.73)
Fiber/disease knowledge	32.99	30.77	30.34
	(23.82)	(24.39)	(24.41)
Cholesterol/disease knowledge	68.15	62.45	64.21
	(26.15)	(31.07)	(29.60)
Total fat/disease knowledge	50.05	45.42	47.93
	(19.85)	(22.86)	(20.83)
Job	.54	.51	.57
	(.49)	(.50)	(.49)
College	.42	.43	.48
	(.49)	(.49)	(.49)
Non-White	.19	.16	.19
	(.39)	(.37)	(.39)
Region 1	.19	.19	.18
	(.39)	(.39)	(.38)
Region 2	.27	.23	.24
	(.44)	(.42)	(.43)
Region 3	.34	.38	.35
	(.47)	(.48)	(.47)
Central	.33	.27	.28
	(.47)	(.44)	(.45)
Suburb	.41	.46	.43
	(.49)	(.49)	(.49)

Table 2. Summary Statistics* (Continued)

Q2	.24	.23	.27
	(.42)	(.42)	(.45)
Q3	.73	.74	.76
	(.44)	(.43)	(.42)
Special Diet	.18	.18	.15
	(.38)	(.38)	(.35)
Household Size	2.65	2.49	2.63
	(1.48)	(1.44)	(1.46)
Age	49.53	54.32	48.64
	(17.26)	(16.98)	(16.82)
Income	33.05	34.36	37.89
	(25.35)	(25.29)	(27.89

*Standard deviations in parentheses.

Table 2. Specification Tests

Equation	Auxiliary R^2	Godfrey/Hutton *J* Tests	Hausman Tests
Fiber	.04	2.44	-1.78*
Cholesterol	.06	23.21*	-2.29*
Total Fat	.03	4.18	-2.16*

*Significant at 5% level.

Table 3. Estimation Results

	Fiber		Cholesterol		Total Fat	
Variables	**OLS**	**IV**	**OLS**	**IV**	**OLS**	**IV**
Constant	17.00 (10.02)	27.56 (4.41)	315.84 (8.19)	491.52 (3.77)	98.34 (12.58)	185.14 (3.61)
Nutrient/Disease Knowledge	.05 (6.65)	-.25 (-1.55)	-.65 (-3.37)	-3.54 (-1.86)	-.07 (-1.58)	-1.83 (-1.73)
Household Size	-.05 (-.36)	-.14 (-.69)	2.83 (.94)	2.40 (.75)	-1.19 (-1.94)	-1.24 (-1.76)
Age	.003 (.19)	.01 (.35)	-.26 (-.84)	-.20 (-.57)	-.29 (-4.69)	-.31 (-3.75)
Income	-.01 (1.27)	-.01 (-1.15)	-.14 (-.73)	-.13 (-.63)	.09 (2.63)	.08 (1.95)
TV	-.05 (-.52)	-.41 (-1.78)	5.63 (2.42)	4.87 (2.07)	.49 (1.36)	-.001 (-.001)
Body Mass Index	.003 (.83)	.003 (.07)	-.70 (-.89)	-.38 (-.45)	.42 (2.25)	.39 (1.88)
Job	.06 (.14)	-.41 (-.57)	14.80 (1.53)	15.72 (1.36)	1.82 (.92)	-1.34 (-.39)
College	.54 (1.33)	1.11 (1.73)	6.26 (.66)	6.76 (.61)	-1.91 (-1.08)	-1.74 (-.67)
Female	.22 (.57)	.01 (.02)	-13.50 (-1.59)	-7.79 (-.71)	-29.34 (-17.79)	-27.77 (-10.71)
Non-White	-.35 (-.68)	-.12 (-.16)	-.48 (-.04)	-7.38 (.49)	-4.16 (-1.85)	-.12 (-.03)
Male/Female Household Head	-2.54 (-3.82)	-1.26 (-1.13)	-33.35 (-2.40)	-13.08 (-.66)	3.40 (1.46)	10.45 (1.83)
Female Household Head	-5.25 (-7.40)	-4.67 (-4.72)	-105.07 (-7.06)	-102.23 (-5.76)	-1.57 (-.59)	.41 (.10)
Smoker	-.58 (-1.40)	-.23 (-.38)	16.57 (1.55)	16.70 (1.38(	6.08 (2.97	5.23 (1.82)

Table 3. Estimation Results (Cont.)

	Fiber		Cholesterol		Total Fat	
Variables	**OLS**	**IV**	**OLS**	**IV**	**OLS**	**IV**
Special Diet	-.02	-.56	-15.00	-17.34	-14.54	-16.37
	(-.05)	(.76)	(-1.35)	(-1.33)	(7.51)	(-5.00)
Program	-.29	.37	-30.37	-40.18	7.09	4.87
	(-.44)	(.38)	(-2.30)	(-2.32)	(2.16)	(1.12)
Disease	-.47	-.27	24.55	27.34	-1.66	-1.18
	(-1.08)	(-.45)	(2.22)	(2.23)	(-.82)	(-.42)
Region 1	-.52	-.68	6.86	.06	-1.16	-3.53
	(-.88)	(-.82)	(.49)	(.00)	(-.45)	(-.87)
Region 2	.14	.97	29.93	24.73	5.98	7.86
	(-.24)	(.99)	(2.31)	(1.67)	(2.36)	(2.14)
Region 3	-.39	.49	24.30	25.63	-1.21	-1.70
	(-.72)	(.55)	(1.99)	1.75	(-.53)	(-.52)
Central	-.26	-.47	19.52	16.71	-4.37	-6.86
	(-.49)	(-.65)	(1.68)	(1.55)	(-1.99)	(-2.02)
Suburb	.14	-.22	26.10	27.47	-3.38	-2.29
	(.29)	(-.31)	(2.50)	(2.28)	(-1.58)	(-.75)
Q1	-.17	.12	8.52	15.74	-4.08	-.99
	(-.32)	(.16)	(.72)	(1.09)	(1.76)	(-.26)
Q2	-.66	-1.17	15.01	11.46	-2.19	-1.68
	(-1.28)	(-1.49)	(1.18)	(.79)	(-.94)	(-.51)
Q3	-.05	-.46	7.75	6.25	-5.25	-5.62
	(-.10)	(-.63)	(.70)	(.48)	(-2.45)	(-1.79)
R^2	.07	.01	.07	.04	.25	.06

Table 4. Out of Sample Comparisons

	1995		1996	
Equation	OLS	IV	OLS	IV
Fiber:				
R^2	.55	.55	.04	.03
Root Mean Square Error	.89	1.00	.56	.57
Mean Percent Square Error	3313	3381	3947	4163
Mean Absolute Deviation	.43	.43	.44	.44
Cholesterol:				
R^2	.01	.00	.01	.004
Root Mean Square Error	.48	.53	1.12	1.18
Mean Percent Square Error	605	685	580	695
Mean Absolute Deviation	.58	.66	.59	.67
Total Fat:				
R^2	.17	.14	.20	.13
Root Mean Square Error	.48	.53	.48	.53
Mean Percent Square Error	210.20	255.20	226	247
Mean Absolute Deviation	.37	.41	.38	.41
Encompassing Test:				
Fiber	.030		-.01	
Cholesterol	-.004		.003	
Total Fat	.010		-.04	