Fitting a Linear Model to Survey Data When the Long-term Average Daily Intake of a Dietary Component Is an Explanatory Variable

Phillip S. Kott, Patricia M. Guenther, David A. Wagstaff, WenYen Juan, and Sibylle Kranz

RTI International, Center for Nutrition Policy and Promotion, Pennsylvania State University, US Food and Drug Administration, and East Carolina University

6110 Executive Blvd.; Rockville, MD 20852/pkott@rti.org

1 Introduction

Much research has been conducted on the effects of dietary intake on health outcomes. One recurring point of criticism in the analysis of diet-disease relationships is the lack of confidence in the measures of the long-term or "usual" intake of dietary components such as energy, nutrients, or food groups. See, for example, Rumpler *et al.* (2008) and Bingham *et al.* (2008).

The quest to improve dietary intake assessment and to achieve more accurate estimates of average daily consumption is ongoing. The primary dietary assessment method used in the National Health and Nutrition Examination Survey, conducted by the United States Center for Disease Control and Prevention, is the 24-hour recall of dietary intake (Moshfegh et al., 2008). Unfortunately, although this method is considered by many to provide the most accurate estimates for individual days, one day of intake is not representative of an individual's longer-term or "usual" diet because of day-to-day variability.

Jackson *et al.* (2008) argue that the average of two 24-hour dietary recalls is not much better and recommend researchers use the average of eight 24-hour recalls as a measure of the usual daily intake of a dietary component. This approach however is not feasible in most studies and certainly not in large, national surveys. Post assessment statistical methods provide promising solutions to this problem. In what follows, we focus on one such method.

Suppose we want to fit a linear regression model relating some outcome variable to a number of explanatory variables, but one or more of those variables are measured with error. As discussed above, this type of measurement error occurs when an individual's long-term average or "usual" daily intake of a dietary component is an explanatory variable in a linear regression model, but that individual's dietary intake is observed on only a small number of days. Using the intake from one of those days (or their average) as a proxy for the individual's usual daily intake can result in badly biased regression estimates. Fortunately, it may be possible to remove nearly all of the bias in this situation by employing *instrumental-variable* regression (see, for example, Fuller, 1987).

To explore this possibility, we use the example of a biomarker of a health-related condition, an individual's long-term (usual) serum beta-carotene concentration, as a function of his or her usual daily intake of beta-carotene. Increased intakes of carotenoids have been found to be associated with decreased risk of chronic diseases, such as age-related macular degeneration, certain type of cancers, and cardiovascular disease (Copper et al. 1999). Of the 34 or so different carotenoids, serum beta-carotene is the one that has been most studied for a potential link to health.

Our goal here is to estimate how that concentration varies among individuals as a function of those individuals' usual dietary intake of beta-carotene. We propose making our estimates using a relatively simple, yet robust, method for which software is readily available.

The National Health and Nutrition Examination Survey (NHANES) is a periodic, stratified, multistage survey. It collects information from nationally representative samples on the health and nutritional status of the non-institutionalized, civilian, US population. NHANES data are released in two-year cycles. The 2003-2004 NHANES data set contains one measure of total serum beta-carotene and beta-carotene intake calculated from each of two 24-hour dietary recalls per sampled individual.

The first dietary recall interview is generally administered on the same day that blood is drawn for the serum beta-carotene measurement. In this study, we consider only food sources of beta-carotene; we do not consider beta-carotene that may be consumed in the form supplements.

We restrict our attention to a subsample of women ages 20-64 from the 2003-2004 NHANES data set. This subsetting to a fairly homogenous population in terms of health facilitates model construction while maintaining a sufficiently large sample size to make estimation practical. We fit a series of linear regressions with the serum beta-carotene concentration as the outcome variable and the first of the one-day beta-carotene intakes as an explanatory variable. Within our instrumental-variable regressions, the second one-day beta-carotene intake serves as the instrumental variable for the first.

Many statistical software packages can conduct instrumental-variable regression. One package, Stata (StataCorp, 2007), which we use here, allows for the incorporation of survey weights into the coefficient-estimation process. Techniques from randomization-based sampling theory are employed by Stata to compute (asymptotic) standard errors when the data comes from a stratified, multistage sample. Little theory exists to interpret what conducting a survey-sensitive instrumental-variable (SSIV) regression is actually doing. An exception is Wu and Fuller (2006). (Humphreys and Skinner, 1997, use instrumental variables to analyze a categorical model with data from a sample survey.)

Kott (2007) discusses two model-driven reasons for using survey weights in a traditional linear regression. We will extend these rationales to SSIV regression, using them to interpret our regressing individual's serum beta-carotene concentration from the 2003-2004 NHANES on the first-day of beta-carotene intake and other explanatory variables.

Section 2 first reviews the theory for fitting a *standard linear model* with instrumental-variable regression when one of the explanatory model variables is measured with unsystematic random error. In the standard model, the expectation of the model error is assumed to be zero when conditioned on the true values of the explanatory variables. The *extended linear model* relaxes that assumption and requires only that the model error be uncorrelated with the true values of the explanatory variables. We show that incorporating the survey weights into an instrumental-variable regression on complex survey data can (asymptotically) remove biases from the coefficient estimates. These biases result either because the standard linear model holds in the population, but not the sample, or because the standard model fails in the population and needs to be replaced by the extended linear model.

Section 3 discusses our subsample of 2003-2004 NHANES data: women ages 20-64 providing two-days of dietary beta-carotene intake data and one measure of serum beta-carotene. The section also describes estimating the variance of an SSIV regression coefficient computed from this subsample.

Section 4 displays the results of regressing the serum beta-carotene concentration of each woman in our subsample on her one-day intake of beta-carotene and other explanatory variables. In order to assess, on the one hand, the impact of using the instrumental variable in this regression and, on the other, the impact of being sensitive to the survey design, linear fits are made with and without the instrumental variable and with and without the survey weights. The standard errors computed for the coefficient estimates of regressions incorporating the survey weights reflect the stratified, multistage design of the NHANES. The regressions ignoring the weights use a more conventional method for computing standard errors. This method ignores not only the survey weights but also the clustering and stratification of the sample.

Section 5 provides a discussion of the strengths and weaknesses of using our proposed SSIV methodology with dietary data from a complex sample survey. Section 6 contains some concluding remarks.

It must be pointed out that our methodology depends on a number of assumptions, some which have been challenged empirically (see, for example, Kipnis *et al.,* 2001). The first is that a single measurement of an individual's serum beta-carotene concentration is an unbiased predictor of his (her) usual concentration. That is less problematic than the following. We assume that an individual's first 24-hour recall provides an unbiased predictor of his (her) usual daily intake of beta-carotene. Although we make no such assumption about the second 24-hour recall, we do assume that the beta-carotene intake from that day is uncorrelated with the measurement error from the first recall (*i.e.,* the difference between the beta-carotene intake from this recall and the individual's usual intake). In making these assumption about 24-hour recalls, we follow an accepted convention in the usual-intake literature which can be found from Nusser *et al.* (1997) to Kipnis *et al*. (2009). The latter goes further than we do and assumes that that a second 24-hour intake provides an unbiased predictor of an individual's usual intake. We remind the reader that all models based on simplifying assumptions are wrong but some are useful.

2 Theory

Although the theory developed in this section can easily be put into a more general form (*e.g.,* with a number of explanatory variables having measurement error), we will express it in the context of the problem at hand: fitting a linear model relating usual serum beta-carotene concentration to usual daily beta-carotene intake and other variables with complex survey data containing one measure of blood concentration and two independent measures of daily intake per sampled individual.

Assume first we have a population of N individuals, and N is very large. Let y_k be the usual serum beta-carotene concentration for individual k, q_k be a single measure of serum beta-carotene concentration for k, x_k be the usual (i.e., long-term average) daily intake of beta-carotene for k (say in a year), p_k be a randomly selected one-day intake of beta-carotene for k, $\mathbf{z}_k$ be the row vector of the $J-1$ additional explanatory variables in the linear model including a constant (or the equivalent), $\mathbf{x}_k = (\mathbf{z}_k \; x_k)$ be the row vector of all J explanatory variables, $\mathbf{p}_k = (\mathbf{z}_k \; p_k)$, h_k be a second randomly selected one-day intake of beta-carotene for k, and $\mathbf{h}_k = (\mathbf{z}_k \; h_k)$.

We will treat q_k, p_k, and h_k as random variables with the following properties:

$$\begin{aligned}
&\sum_{k=1}^{N} (q_k - y_k)/N = \mathrm{O}_P(1/\sqrt{N}),\\
&\sum_{k=1}^{N} (p_k - x_k)/N = \mathrm{O}_P(1/\sqrt{N}),\\
&\sum_{k=1}^{N} h_k(q_k - y_k)/N = \mathrm{O}_P(1/\sqrt{N}), \qquad (1)\\
&\sum_{k=1}^{N} h_k(p_k - x_k)/N = \mathrm{O}_P(1/\sqrt{N}), \text{ and}\\
&\sum_{k=1}^{N} \mathbf{h}_k'\mathbf{x}_k/N - \mathbf{\Psi} = \mathbf{O}_P(1/\sqrt{N}),
\end{aligned}$$

for some matrix $\mathbf{\Psi}$ having full rank. The assumptions in equation (1) mean that the biases in *measurement errors* in q_k and p_k as predictors of y_k and x_k respectively (*i.e.,* $q_k - y_k$ and $p_k - x_k$), are asymptotically zero. Moreover, h_k is asymptotically uncorrelated with the measurement errors of q_k and p_k.

Assuming the measurement errors in q_k and p_k are asymptotically zero means there are no systematic biases in the data-collecting instruments. Observe that we are *not* assuming that the measurement error in h_k is also asymptotically zero.

Under the assumptions in equation (1), the vector

$$\mathbf{B}_{IV} = \left(\sum_{k=1}^{N} \mathbf{h}_k'\mathbf{p}_k\right)^{-1} \sum_{k=1}^{N} \mathbf{h}_k' q_k \qquad (2)$$

is a consistent estimator for the parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)'$ in the *standard linear model* (with an instrumental variable):

$$y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k,$$
$$\text{where } E(\varepsilon_k|\, \mathbf{x}_k, h_k) = 0, \tag{3}$$

and $\sum^N \mathbf{h}_k'\varepsilon_k /N = \mathbf{O}_P(1/\sqrt{N})$. The proof of a variant of this assertion can be found in Fuller (1987). We have not included formal proofs in this paper because they are either well known or trivial extensions of others in the literature. For our purposes, it is helpful to note that $\mathbf{B}_{IV} - \boldsymbol{\beta} = (\sum^N \mathbf{h}_k'\mathbf{p}_k)^{-1}(\sum^N \mathbf{h}_k'[y_k - \mathbf{p}_k\boldsymbol{\beta}]) \approx \boldsymbol{\Psi}^{-1} \sum^N \mathbf{u}_k /N$, where

$$\begin{aligned}\mathbf{u}_k &= \mathbf{h}_k'(y_k - \mathbf{p}_k\boldsymbol{\beta}) \\ &= \mathbf{h}_k'[\varepsilon_k + (q_k - y_k) - (\mathbf{p}_k - \mathbf{x}_k)\boldsymbol{\beta}],\end{aligned} \tag{4}$$

and $\sum^N \mathbf{u}_k /N = \mathbf{O}_P(1/\sqrt{N})$. Note that $\mathbf{u}_k$ can be viewed as the instrumental-vector-scaled element error, $\mathbf{u}_k = \mathbf{h}_k'e_k$, where $e_k = y_k - \mathbf{p}_k\boldsymbol{\beta}$ is the element error (more on this value in the following section).

The only difference between the model in equation (3) and the usual standard linear regression model is that the expectation of the error term ε_k is zero conditioned on the instrumental variable as well as the explanatory variables. This expansion of the usual standard model is very mild compared to the assumptions in equation (1).

Unfortunately, models can fail. That is frequently the case when one tries to fit sample survey data to a linear model because of the limited number of potential explanatory variables available from the survey. Kott (2007) proposed an *extended linear model* that, in the absence of the need for instrumental variables, will almost always hold. In our context, it is

$$y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k, \tag{5}$$
$$\text{where } E(\mathbf{x}_k\varepsilon_k) = E(h_k\varepsilon_k) = 0.$$

Of course, if we had full information on all N individuals in our population, it would be more sensible to estimate $\boldsymbol{\beta}$ with $\mathbf{B}_{OLS} = (\sum^N \mathbf{x}_k'\mathbf{x}_k)^{-1} \sum^N \mathbf{x}_k'y_k$ than with $\mathbf{B}_{IV}$. Often, however, we only have partial information from a sample of m individuals. In our case, the sampled individuals have been selected using a stratified, multistage survey subject to nonresponse. Moreover, only one serum measurement and two independent one-day beta-carotene intakes have been collected from each sampled individual.

To estimate model parameters like $\boldsymbol{\beta}$, the NHANES provides individual survey weights. We will assume that the survey weights, w_k, are such that

$$\begin{aligned}\sum_{k=1}^{N} w_k I_k(q_k - y_k)/N &= O_P(1/\sqrt{n}), \\ \sum_{k=1}^{N} w_k I_k(p_k - x_k)/N &= O_P(1/\sqrt{n}), \\ \sum_{k=1}^{N} w_k I_k h_k(q_k - y_k)/N &= O_P(1\sqrt{/n}), \\ \sum_{k=1}^{N} w_k I_k h_k(p_k - x_k)/N &= O_P(1/\sqrt{n}),\end{aligned} \tag{6}$$

$$\sum_{k=1}^{N} w_k I_k \mathbf{h}_k' \boldsymbol{\varepsilon}_k / N = \mathbf{O}_P(1/\sqrt{n}), \text{ and}$$

$$\sum_{k=1}^{N} w_k I_k \, \mathbf{h}_k' \mathbf{x}_k / \mathrm{N} \; - \; \boldsymbol{\Psi} = \mathbf{O}_P(1/\sqrt{n}),$$

where $\boldsymbol{\Psi}$ again has full rank. In equation (6), I_k is an indicator variable equal to 1 when k is in the (respondent) sample and 0 otherwise, and n is the number of primary sampling units (PSUs) selected for the sample, which we assume to be large.

Accepting the assumptions in equation (6),

$$\mathbf{b}_{IV} = \left(\sum_{k=1}^{N} w_k I_k \mathbf{h}_k' \mathbf{p}_k \right)^{-1} \sum_{k=1}^{N} w_k I_k \mathbf{h}_k' q_k, \tag{7}$$

which can be computed with data from a sample survey, is a consistent estimator for $\boldsymbol{\beta}$ under either the standard model in equation (3) or the extended model in equation (4). It is important to realize that the standard linear model allows the possibility, $E(\varepsilon_k | \mathbf{x}_k, h_k, I_k) \neq 0$; that is to say, the standard model may fit in the population but not in the sample. This can happen when the model errors, the ε_k, are correlated with the individual probabilities of sample selection. In other words, when the surveys weights are not ignorable with respect to the model, they need to be incorporated into the estimated instrumental-variable regression coefficient $\mathbf{b}_{IV}$.

3 The NHANES Data and Variance Estimation

We focus here on a 2003-2004 NHANES data set containing health and nutritional information from sampled individuals providing *two* nonconsecutive (*i.e.*, independent) 24-hour dietary intakes. Each individual in that data set was assigned a survey weight roughly equal to the inverse of the individual's probability of selection and response into the data set. The weighting procedures also attempt to balance the days of the week. This was complicated by the nature of the two-intake-day data set. See NCHS (2007) for more details on the weighting process.

The data set also contains indicators of the (pseudo) primary sampling unit (PSUs) and first-stage (pseudo) stratum for each sampled individual for variance estimation purpose; the actual PSUs and strata are masked for confidentiality reasons. There are two sampled PSUs in each of the 15 strata in the 2003-2004 NHANES data set.

Consider a particular subsample of the NHANES data set to be used for estimating parameters of a particular target population, in our case civilian, non-institutionalized women, ages 20-64, in the 2003-2004 US population. Invoking the notation from the previous section, let $\mathbf{u}_{gi} = \sum w_k I_k \mathbf{u}_k$, where the summation is over every individual in the population assigned to PSU i of stratum g, and the $\mathbf{u}_k = \mathbf{h}_k' e_k$ are the instrumental-vector-scaled element errors defined in equation (4).

We assume for variance estimation purposes that the $\mathbf{u}_{gi}$ are independent random variables and the two $\mathbf{u}_{gi}$ within each stratum have a common mean. Note that when the $y_k | \mathbf{x}_k$ are identically distributed regardless of stratum, this mean is zero for all $\mathbf{u}_{gi}$. By *not* assuming the $\mathbf{u}_{gi}$ have mean zero, we are allowing the possibility that the stratification in the sample design matters. By aggregating the instrumental-vector-scaled errors to the PSU, we are allowing for the possibility that the element errors with each PSU are correlated.

A nearly (*i.e.*, asymptotically) unbiased estimator for the variance of $\mathbf{b}_{IV}$ under the extended model (and also the stronger standard model) has the form,

$$\hat{\mathbf{V}} = \mathbf{ACA}', \tag{8}$$

where $\mathbf{A} = (\sum^{N} w_k I_k \mathbf{h}_k' \mathbf{p}_k)^{-1}$, $\mathbf{C} = \sum^{15} (\hat{\mathbf{u}}_{g1} - \hat{\mathbf{u}}_{g2})(\hat{\mathbf{u}}_{g1} - \hat{\mathbf{u}}_{g2})'$, $\hat{\mathbf{u}}_{gi} = \sum w_k I_k \mathbf{h}_k' (q_k - \mathbf{p}_k \mathbf{b}_{IV})$, and (again) the summation is over every individual in the population assigned to PSU *i* of stratum *g*. This is what Stata computes.

More generally, when there are n_g sampled PSUs in stratum *g* out of *G* strata,

$$\mathbf{C} = \sum^{G} [n_g/(n_g - 1)] \{ \sum^{n_g} \hat{\mathbf{u}}_{gi} \hat{\mathbf{u}}_{gi}' - (\sum^{n_g} \hat{\mathbf{u}}_{gi})(\sum^{n_g} \hat{\mathbf{u}}_{gi})'/n_h \}.$$

When there is only one stratum, $\mathbf{C} = \sum^{n} [n/(n-1)] \sum^{n} \hat{\mathbf{u}}_{1i} \hat{\mathbf{u}}_{1i}'$ (because $\sum^{n} \hat{\mathbf{u}}_{1i} = \mathbf{0}$), and $\hat{\mathbf{V}} = \mathbf{ACA}'$ has the form of a robust variance estimator with a clustered sample.

Under simple random sampling, where $m = n$ and the w_k are all 1, $\mathbf{C}$ becomes $\mathbf{C}_R = \sum^{N} [m/(m-1)] \sum^{N} I_k \hat{\mathbf{u}}_k \hat{\mathbf{u}}_k'$, and

$$\hat{\mathbf{V}}_R = \mathbf{AC}_R\mathbf{A}'. \tag{9}$$

In contrast, the conventional variance estimator not only ignores the sampling design (weights, clustering, and stratification) and it assumes the element errors, the $e_k = \varepsilon_k + (q_k - y_k) - (p_k - x_k)\beta_J$, are homoscedastic. The conventional variance estimator is

$$\hat{\mathbf{V}}_S = (\sum_{k=1}^{N} I_k \mathbf{h}_k' \mathbf{p}_k) s^2 \tag{10}$$

where $s^2 = \sum^{N} I_k (q_k - \mathbf{p}_k \mathbf{b}_{IV})^2/(m-J)$.

The element error, e_k, has three components: the model error (ε_k), the error of the observed measure of serum beta-carotene concentration as a predictor for usual serum beta-carotene concentration ($q_k - y_k$), and a fixed multiple of the error of the one-day beta-carotene intake as a predictor for usual intake of beta-carotene ($(p_k - x_k)\beta_J$); recall that only one member of $\mathbf{p}_k - \mathbf{x}_k$, namely $p_k - x_k$, is nonzero in our framework). Although it is not necessary that all three components be homoscedastic for e_k to be homoscedastic, it is unlikely that the components' heteroscedasticity perfectly counterbalance each other. As a consequence, even when the sample design is ignorable, and the instrumental-variance regression coefficient $\mathbf{b}_{IV}$ in equation (7) is estimated setting all the w_k to 1, it may be prudent to use the robust-to-unspecified-heteroscedasticity variance estimator $\hat{\mathbf{V}}_R$ rather than $\hat{\mathbf{V}}_S$.

The standard error for a component of $\mathbf{b}_{IV}$ is the square root of the corresponding diagonal of $\hat{\mathbf{V}}$. Since $\hat{\mathbf{V}}$ is only nearly unbiased, standard errors of regression coefficients computed using randomization-based techniques are sometimes called "asymptotic standard errors." The same nomenclature applies to standard errors from instrumental-variable regression based on $\hat{\mathbf{V}}_S$ even when the sampling design can be ignored and the

$$e_k = \varepsilon_k + (q_k - y_k) - (p_k - x_k)\beta_J$$

are homoscedastic because the estimated regression coefficient itself is not unbiased, only consistent.

We confine our analyses in the next section to the 1,317 women ages 20 to 64 in the 2003-2004 NHANES who provided two days of dietary intake and a measure of serum beta-carotene concentration. (For reasons explained in the next section, one of these women is later dropped from our analytic data set.) We adjusted the two-day survey weight for each woman in our subsample by scaling her NCHS-supplied two-day dietary weight to the sum of these weights across all the 20-64 year-old women in the sampled PSU, divided by the sum restricted to women in our subsample (*i.e.*, those with both two days of dietary intake and one serum beta-carotene measurement). This effectively assumes that serum beta-carotene measures are missing at random within PSUs.

4 The Results

The upper left-hand corner of Table 1 displays the coefficient estimates from naïvely using ordinary least squares to regress the one measure of serum beta-carotene concentration (measured in micrograms per deciliter, ug/dL) for each woman in our subsample on her first day of beta-carotene intake (measured in milligrams, mg) and other explanatory variables. To conduct ordinary least squares, the $\mathbf{h}_k$ and w_k in equation (5) were changed to $\mathbf{p}_k$ and 1, respectively.

In addition to first-day beta-carotene intake (*Intake*), the other explanatory variables in what we label the "original model" are the woman's age in years centered at 43 (*Age*), an indicator of whether the woman was both a current cigarette smoker and had smoked at least 100 cigarettes in her life (*Smoker*), an indicator of whether the woman was of Hispanic origin (*Hispanic*), and an indicator of whether the woman's family income was at or above 185% of the Federal poverty threshold (*Poverty Income Ratio* or PIR ≥ 185%) and an indicator of whether it was above 350% of the Federal poverty threshold (PIR > 350%). See US Bureau of the Census (2007) for a description of how the Federal poverty threshold is computed. A family income of less than 185% of the Federal poverty threshold is required to qualify for the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). A family with a Poverty Income Ratio in excess of 350% – the maximum level at which there is any Federal or state nutrition and health-insurance assistance program available – is considered middle to high income.

The inclusion of each of these explanatory variables decreases the model root mean squared error of the regression fit, and each is available for all of the women in our subsample set except one. Including an indicator for Hispanic-origin provides a slightly better fit than including an indicator for Mexican-American. Including a race variable does not improve the fit nor does incorporating other versions of the age variable.

Current cigarette-smoking status was only queried in the NHANES of participants who had smoked at least 100 cigarettes in their entire lifetime. One of our 1,317 women did not answer whether she had smoked at least 100 cigarettes and is, therefore, excluded from the regression analyses. Since this is less than 0.1% of the subsample, little harm results. We do not include explanatory variables on cigar- and pipe-smokers in our analyses because a greater percentage of the respective answers were missing from the 1,317 women in our subsample.

Standard errors in the ordinary-least-squares regression are computed as if a standard model relating serum beta-carotene concentration to first-day beta-carotene intake and the other explanatory variables is correctly specified, the attributes of the sample design can be ignored, and the element model errors are uncorrelated and homoscedastic. These assumptions appear to be wrong as evidenced by the very-different coefficient estimates resulting from an analogous survey-weighted-least-squares regression to the right of the initial fit in table.

The weighted-least-squared regression assumes only that the extended model relating beta-carotene concentration to one-day beta-carotene intake and the other explanatory variables is correct, an assumption that holds whenever regression estimates can be computed. Standard errors are computed using the survey sensitive routine in Stata, and the resulting "*t*-values" used in calculating the P[robability]>|*t*| column have 15 degrees of freedom: the number of PSUs minus the number of strata.

The P>|*t*| column relies on commonly made, but dubious, assumptions that the $\hat{\mathbf{u}}_{gi}$ are nearly normal and have roughly equal variances. We will not use it in our model assessments, nor will we assume that the *t*-values are anything other than the standard errors of the estimated coefficients divided by the coefficients themselves. Under the null hypothesis that the true value of a coefficient is zero, we expect the square of the associated *t*-value to be approximately equal to 1.

Judging from the weighted-least-squares estimates in the second set of results in Table 1, we might choose to remove the Hispanic-origin indicator from the model since the associated *t*-value is less than 1 in absolute value. Determining explanatory variables using the weighted-regression fit and an absolute-*t*-value-greater-than-1 rule does not change any of our other explanatory-variable determinations (*e.g.,* the candidate race variable is still excluded).

The instrumental-variable regressions on the right-hand side of Table 1 use the second day of beta-carotene intake as an instrumental variable for the first-day intake. One is unweighted; the other is weighted.

It is well known that an individual's intake of a nutrient often varies by the day of the week (see, for example, Nusser *et al*.,

1996). NCHS weights the NHANES dietary data sets with this in mind. Moreover, NHANES protocol calls for collecting the second dietary recall for a sampled individual on a different day of the week from the first. For our purposes, however, this effort has the potential for inducing a negative correlation between the two daily intakes recorded for the same individual, which would violate one of the assumptions of instrumental-variance regression.

In order to remove this potential for negative correlation, we adjust the first-day beta-carotene intakes from food in our weighted instrumental-variable analyses to remove the day-of-the-week effect. Paralleling a multiplicative adjustment in Nusser *et al.* (1996), we replace each p_k with

$$p_k^a = \frac{\frac{1}{7}\sum_{d=1}^{7} \frac{\sum_{f \in S_d} w_f p_f}{\sum_{f \in S_d} w_f}}{\frac{\sum_{i \in S_d} w_i p_i}{\sum_{i \in S_d} w_i}} p_k , \tag{11}$$

where the w_f are the weights described in the last section, and S_d is the subset of our 1,316 women whose first day of beta-carotene intake was on the day of the week denoted by d, where $d = 1$ was Sunday, $d = 2$ Monday, and so forth. Observe that the weighted-mean of the p_k^a for each day of the week is identical.

Coefficient standard errors for the unweighted instrumental-variable regression are computed using $\hat{\mathbf{V}}_S$ in equation (10), while coefficient standard errors for weighed instrumental-variable regression use $\hat{\mathbf{V}}$ in equation (8).

The unweighted fit in the instrumental-variable regression is based on a more scientifically plausible standard model than the ordinary-least-squares regression since it relates to serum beta-carotene concentration (both a single measure and the usual concentration) to usual daily intake of beta-carotene and other explanatory variables. Still, it appears that incorporating weights into the instrumental-variable regression may be needed to avoid bias in the coefficient estimates for age and the family income indicators.

The most important result comes from comparing the least-squares regression fits with the instrumental-variable fits. The coefficient on beta-carotene intake is more than four times greater using instrumental-variable regression. This is because serum beta-carotene concentration is a function of *usual* beta-carotene intake, not one-day beta-carotene intake, even though the daily intake value in the least-squared regression was generally for the day before the blood concentration was measured.

Although it is possible for the standard model relating serum beta-carotene concentration to usual daily beta-carotene intake and the other explanatory variables to hold, it appears some of the coefficients are biased when the instrumental-variable regression is unweighted. In fact, the SSIV regression in Table 1 suggests that not only the Hispanic-origin indicator but also the PIR≥350% indicator can be removed from the model. This is not clear in the unweighted IV regression even though the estimated standard error for this variable is about the same in both regressions (around 2.1).

The Hispanic-origin and PIR≥350% indicators are removed from the model estimated in Table 2. The other family-income indicator is redefined to keep the intercept significant. (We drop the Hispanic-indicator from this model even though its absolute t value in the SSIV fit of the original model is greater than 1 because its sign is opposite of what we had been led to expect from the previous regressions.)

The fits of the "final model" in Table 2 tell essentially the same story as Table 1: weights may matter only a little in this context (although using weights helped us see that the PIR≥350% indicator probably did not belong in the model), but using instrumental-variable regression to determine the relationship between the biomarker in blood and the usual dietary intake matters a great deal.

5 Discussion

Our survey-sensitive instrumental-variable (SSIV) regression methodology allows us to estimate the impact of a change in the usual daily intake of a dietary component for an individual on a biomarker for health-related outcome for that individual using survey data containing only two independent 24-hour dietary recalls. To do this, we must assume that the reported intakes for the first of the 24-hour recalls, after adjusting for day-of-the-week effect (see equation (11)), are free of systematic reporting error (i.e., are unbiased predictors of the individuals' usual daily intakes) and that the survey weights accurately reflect the probabilities of individuals' selection into and response to the survey.

In Table 2, we saw how treating one 24-hour recall as a proxy for usual daily intake appears to seriously bias the modeling of the relationship between usual serum beta-carotene and usual daily intake of beta-carotene, age, smoking status, and family income among women ages 20-64. Such treatment of one 24-hour recall is not uncommon in the nutritional literature. See, for example, Grandjean *et al.* (2008). Modest biases were also caused by ignoring the survey weights associated with the survey data.

Table 2 also revealed some of the price we may have to pay for using the SSIV method: increased standard errors. The increases in standard error due to using an instrumental-variable were very large, from under 0.5 to nearly 2.0 for the intake coefficient in the weighted regressions. The increases due to incorporating the survey weights were less profound, but still noticeable. When the weights were ignored, the standard error for the intake coefficient in the instrumental-variable regression was less than 1.6.

Table 3 shows that the impact of weighting on standard error had more to do with the assumption that the combination of element errors are homoscedastic than with incorporating the weights or with being sensitive to the clustering and stratification in the sample. Using equation (9) to compute a more robust set of unweighted standard errors leads to a standard error for the intake coefficient under instrumental-variable regression of slightly more than 2.0 (as opposed to less than 1.6).

Table 3 also shows that using the average of the two 24-recalls as a proxy for usual intake apparently removes some of the bias in the intake coefficient but not much. The unweighted coefficient increases from roughly 1.1 when using only one recall to 1.8 (the weighted coefficient estimates are similar but not shown), which is still far short of approximately 6, the coefficient estimate under unweighted instrumental-variable regression.

Surprisingly, the estimated mean intake of beta-carotene on the second day among our 1,317 women was slightly larger than the first; however, this difference was not significant (at the 0.05 level). It is tempting to argue that the increase in standard error due to using instrumental-variable regression is likewise chimerical. The *t*-value for the intake coefficient in the weighted regression in Table 2 is 2.9, while the analogous *t*-value for the SSIV regression is only slightly higher, 3.0. Viewing a *t*-value as a normalized standard error, these numbers appear to be nearly identical.

This argument does not mitigate the fact that the errors in the SSIV regression coefficients come from three sources, the model error in equation (3) (ε_k), the measurement error in the serum beta-carotene concentration ($q_k - y_k$), and the measurement error in the day-of-the-week-adjusted first-day intake ($p_k^a - x_k$). There is nothing we can do about the first except draw a larger sample. The second is likely to be small in this context (as shown in Lin *et al.* 1998). The size of the third, however, reflects our using the information from the second day of intake only as an instrument rather than making it part usual intake measure itself.

Regression calibration is a technique that allows a more efficient use of the second day of intake data while also potentially removing the bias in regression modeling that employing a simple two-day average would cause. Kipnis *et al.* (2009) describe how to use regression calibration when the explanatory dietary-intake variable in the model is an infrequently consumed food, making the variance of p_k^a particularly large.

Regression calibration can also overcome another shortcoming of our approach to instrumental-variable regression in this context. The model relating the biomarker of a health outcome to the usual daily dietary-intake of a dietary component need not be linear. There is a steep price to be paid, however. Regression calibration requires the assumption of a tightly specified

stochastic model for daily dietary intakes across the individuals in the target population in order to compute an optimal proxy for an individual's usual intake. As described in Kipnis *et al*., the appropriate modeling of individual daily intakes in order to determine the optimal proxy is not a trivial exercise even when the dietary component (food or nutrient) is consumed every day. Moreover, the complex model(s) the authors develop ignores the clustering and stratification in the sampling design (the model has two parts when the dietary component is not consumed every day).

Our SSIV regression fits a simple linear model relating a biomarker of a heath-related outcome (in our case, serum beta-carotene) and a usual daily dietary-intake variable (beta-carotene), which some may view as a limitation. Assuming the extended model, however, provides protection against model failure. In the standard model, model failure can take of form of a misspecified functional form or missing explanatory variables. The extended model all but removes the possibility of misspecification if our assumptions about the relationship among one-day intakes based on 24-hour recalls and usual daily intake and between a single outcome measurement its longer-term average value are correct. The interpretation of the results changes slightly, however.

We can see how by returning to the last regression in Table 2. Given the age of a woman from 20 to 64 in the 2003-2004 non-institutionalized, civilian population, her smoking status, and whether or not the her family income was less than 185% of the Federal poverty threshold, a one milligram increase in her usual daily beta-carotene intake results in an estimated 5.8789 microgram per deciliter increase in her serum beta-carotene concentration. Under the standard model, this impact on serum beta-carotene is the same, plus or minus a random error associated with the individual, *no matter what the usual daily intake of the beta-carotene*. Under the extended model, this need no longer be the case. The estimated impact of a one milligram increase of usual daily beta-carotene intake on serum beta-carotene, holding the other explanatory variables constant, is 5.9 micrograms per deciliter *on average* across all possible usual intakes.

No similar distinction exists between the interpretations of the standard and extended models for the smoking and family-income explanatory variables because they are categorical. There would be such a distinction for the age variable, in principle, except that our model fitting rejected the inclusion of the additional explanatory variable age^2, suggesting that the impact of age on serum beta-carotene, holding the other explanatory variables constant, is as specified.

6 Concluding Remarks

We have proposed a survey-sensitive instrumental-variable method for conducting a linear regression relating a biomarker for a health-related outcome measure in a complex survey like the NHANES to explanatory variables including the usual daily intake of a dietary component despite the limitation of the NHANES collecting no more than two 24-hour dietary recalls per sampled individual. We have shown that our method returns nearly unbiased coefficient estimates under certain conditions which are numerous and subject to question, but relatively mild for dietary studies. The methodology can be conducted with readily available software and extends easily to the incorporation of more than one dietary component. The extension itself is left to the reader.

Some may question a model of serum beta-carotene concentration that does not include dietary supplements, others our use of a strictly linear model. Neither criticism undermines the methodology itself. Moreover, one of the advantages of our simple linear approach is that it allows an extended-model interpretation that does not require the model to be as complete and free of misspecification error as do more complicated statistical methods. Under neither the extended nor standard-model interpretations do the error components (ε_k, $(q_k - y_k)$, and $(p_k - x_k)\beta_J$) need to have particular distributions.

There are competitors to our SSIV methodology. One such method is regression calibration. It too produces nearly unbiased parameter estimates under certain conditions. It is more difficult to execute, however, requiring the user to conduct a classical measurment-error analysis for which the sample data must be transformed so that the components of the analysis can be assumed homoscedastic and normally distributed. Furthermore, the user of regression calibration needs to make even more questionable assumptions than with our linear SSIV methodology (so far as we know, the complex sampling structure must ignored in the measurement-error modeling step of regression calibration).

When the user is willing to make these additional assumptions and overcome the technical difficulties, the resulting coefficient will likely be more efficient than those returned by SSIV regression. We believe, however, that making unverifiable assumptions should be done as sparingly as possible. For many scientific purposes, the use of our more robust methodology should be considered before its competitors.

Acknowledgements

The authors thank Wayne Fuller, who suggested to two of us that instrumental-variable regression could be used in this context fifteen years ago.

References

Bingham, S., Luben, R. Welch, A., Low, Y.L., Khaw, K.T. Wareham, N., & Day, N. (2008). Associations Between Dietary Methods and Biomarkers, and Between Fruits and Vegetables and Risk of Ischaemic Heart Disease, in the EPIC Norfolk Cohort Study. *International Journal of Epidemiology*, *37(5)*, 978 - 987.

Copper, D.A., Eldridge, A.L., & Peters, J.C. (1999). Dietary Carotenoids and Certain Cancers, Heart Disease, and Age-related Macular Degeneration: A Review of Recent Research. *Nutrition Reviews*, *57*, 201-214.

Fuller, W.A. (1987). *Measurement Error Models*, New York: Wiley.

Grandjean, A.C., Fulgoni, V.L. Reimers K. J. , & Agarwal S. (2008). Popcorn Consumption and Dietary Physiological Parameters of US Children and Adults: Analysis of the Health and Nutrition Examination Survey (NHANES) 1999-2002 Dietary Survey Data. *Journal of the American Dietetic Association*, *108(5),* 853-856.

Humphreys, K. & Skinner, C.J. (1997) Instrumental Variable Estimation of Gross Flows in the Presence of Measurement Error. *Survey Methodology*, *23,* 53-60.

Jackson, K.A., Byrne, N.M., Magarey, A.M. , & Hills A.P. (2008). Minimizing Random Error in Dietary Intakes Assessed by 24-hour Recall in Overweight and Obese Adults. *European Journal of Clinical Nutrition*, *62*, 537-543.

Kipnis, V., Midthune, D., Buckman, D.W., Dodd, K. W. Guenther, P.M. Krebs-Smith, S.M. Subar, A.F., Tooze, J.A., Carroll. R.J. & Freedman, L. S. (2009). Modeling Data with Excess Zeros and Measurement Error: Application to Evaluating Relationships Between Episodically Consumed Foods and Health Outcomes. *Biometrics*, forthcoming.

Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Schatzkin, A., Subar, A.F. &, Carroll. R.J. (2001). Empirical Evidence of Correlated Biases in Dietary Assessment Instruments and Its Implications. *American Journal of Epidemiology*, 153, 394-403.

Lin, Y,, Burri, B.J., Neidlinger, T.R., Muller, H.G., Dueker, S.R., & Clifford, A. (1998). Estimating the Concentration of Beta-carotene Required for Maximal Protection of Low-density Lipoproteins in Women. *American Journal of Clinical Nutrition, 67*, 837-845.

Kott, P.S. (2007). Clarifying Some Issues in the Regression Analysis of Survey Data. *Survey Research Methods*, 1, 11-18.

Moshfegh, A.J., Rhodes, D.G., Baer, D.J., Murayi, T., Clemens, J.C., Rumpler, W.V., Paul, D.R., Sebastian, R.S., Kuczynski, K.J., Ingwersen, L.A., Staples, R.C., & Cleveland, L.E. (2008). The US Department of Agriculture Automated Multiple-Pass Method Reduces Bias in the Collection of Energy Intakes. *American Journal of Clinical Nutrition*, *88*, 324-332.

National Center for Health Statistics (2007). *National Health and Examination Survey 2003 -2004, Documentation, Codebook, and Frequencies: Dietary Interview - Total Nutrient Intakes (Second Day)*. (US Department of Health of Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics; Available online at http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/dr2tot_c.pdf.)

Nusser, S.M., Carriquiry, A.L., Dodd, K.W. & Fuller, W.A. (1996). A Semi-parametric Transformation Approach to Estimating Usual Nutrient Intake Distributions,. *Journal of the American Statistical Association, 91*, 1440-1449.

Rumpler, W.V., Rhodes, D.G., Moshfegh, A.J., Paul, D.R., Kramer, M.H. (2008). Identifying Sources of Reporting Error Using Measured Food Intake. *European Journal of Clinical Nutrition*, *62(4)*, 544-52.

Statacorp (2007). *Stata Statistical Software: Release 10*. (StataCorp LP, College Station, TX.)

US Bureau of the Census (2008). *How the Census Bureau Measures Poverty.* (US Department of Commerce, Bureau of the Census; Available online at: http://www.census.gov/hhes/www/poverty/povdef.html.)

Wu, Y.Y. & Fuller, W.A. (2006). *Estimation of Regression Coefficients with Unequal Probability Samples*. (*Proceedings of the Survey Research Methods Section*, American Statistical Association, 3892-3899.)

Table 1. Linear regression of original model for serum beta-carotene (*ug/dL*).

	Unweighted Least Squares				Weighted Least Squares				Unweighted IV Regression				Weighted IV Regression			
	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|
Intercept	15.0891	0.9496	15.89	0.000	15.2714	1.5380	9.93	0.000	5.9627	3.2013	1.86	0.063	8.2098	3.1152	2.64	0.019
Intake	1.1174	0.1230	9.08	0.000	1.3543	0.4665	2.90	0.011	6.0668	1.5677	3.87	0.000	5.7904	1.9231	3.01	0.009
Age	0.1001	0.0373	2.69	0.007	0.1424	0.0553	2.57	0.021	0.0658	0.0566	1.16	0.245	0.1023	0.0522	1.96	0.069
Smoker	-7.6596	1.1700	-6.55	0.000	-8.0473	1.2465	-6.46	0.000	-7.1685	1.7520	-4.09	0.000	-7.1030	2.1153	-3.36	0.004
Hispanic	1.7031	1.1625	1.47	0.143	0.4436	1.7420	0.25	0.802	0.9985	1.7581	0.57	0.568	-3.2023	2.2047	-1.45	0.167
PIR≥185%	1.8593	1.2267	1.52	0.130	3.3376	1.5377	2.17	0.046	3.2024	1.8779	1.71	0.088	3.5505	1.8329	1.94	0.072
PIR>350%	5.4012	1.3074	4.13	0.000	3.4448	1.7245	2.00	0.064	2.7948	2.1154	1.32	0.186	1.1725	2.0703	0.57	0.580

Intake is the one-day intake of beta-carotene in food measured in *mg*. For the last regression, one-day intake is day-of-the-week adjusted.

SE is Standard error. In the unweighted regressions, these estimates assume the survey design is ignorable and the combined errors are homoscedastic. In the weighted regressions, they incorporate the survey design as described in text.

Age is age in years minus 43.

Smoker is a current cigarette smoker who has smoked at least 100 cigarettes in her lifetime.

PIR (poverty income ratio) is the ratio of family income to the Federal poverty threshold, expressed as a percentage.

Table 2. Linear regression of final model for serum beta-carotene (*ug/dL*).

	Unweighted Least Squares				Weighted Least Squares				Unweighted IV Regression				Weighted IV Regression			
	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|
Intercept	20.3798	0.7214	28.25	0.000	20.7888	1.5598	13.33	0.000	10.8609	3.1475	3.45	0.001	12.0327	4.0327	2.98	0.009
Intake	1.1437	0.1236	9.25	0.000	1.3782	0.4746	2.90	0.011	6.1342	1.5612	3.93	0.000	5.8789	1.9559	3.01	0.009
Age	0.1013	0.0375	2.70	0.007	0.1372	0.0570	2.41	0.029	0.0661	0.0571	1.16	0.247	0.1068	0.0517	2.07	0.057
Smoker	-8.3628	1.1596	-7.21	0.000	-8.5176	1.1546	-7.38	0.000	-7.5386	1.7524	-4.30	0.000	-6.9852	2.1343	-3.27	0.005
PIR<185%	-4.5599	0.9918	-4.60	0.000	-5.3194	1.3512	-3.94	0.001	-4.5795	1.4827	-3.09	0.002	-4.6263	1.4212	-3.26	0.005

See Table 1 for definitions.

Table 3. Three robust unweighted linear regressions of final model for serum beta-carotene (*ug/dL*) compared to weighted IV regression.

	First Day as Intake				Two-day Average as Intake				Unweighted IV Regression				Weighted IV Regression			
	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|	Estimate	SE	*t*-value	P>|*t*|
Intercept	20.3798	1.0392	19.61	0.000	18.5631	0.9727	19.08	0.000	10.8609	3.8298	2.84	0.005	12.0327	4.0327	2.98	0.009
Intake	1.1437	0.3872	2.95	0.003	1.7792	0.3402	5.23	0.000	6.1342	2.0327	3.02	0.003	5.8789	1.9559	3.01	0.009
Age	0.1013	0.0336	3.01	0.003	0.0860	0.0332	2.59	0.010	0.0661	0.0470	1.16	0.160	0.1068	0.0517	2.07	0.057
Smoker	-8.3628	0.8675	-9.64	0.000	-7.7006	0.8624	-8.93	0.000	-7.5386	1.9243	-4.30	0.000	-6.9852	2.1343	-3.27	0.005
PIR<185%	-4.5599	0.9279	-4.91	0.000	-4.1836	0.9727	-4.50	0.000	-4.5795	1.4071	-3.09	0.001	-4.6263	1.4212	-3.26	0.005

See Table 1 for definitions; except

Intake is a proxy for usual daily intake, which is defined in the header for the first two regressions.

SE (standard error) is computed in the unweighted regressions in a manner robust to unspecified heteroscedasticity.