Measurement Error Models: Applications to Nutrition Data

Brenna Curley

Iowa State University

Grinnell College Seminar
February 24, 2014
Motivation

What given:
- National surveys only collect a limited number of days of nutritional data for each individual.

But:
- Longterm average intakes of nutrients are of interest.

Solution:
- Minimize nuisance day-to-day variation using a measurement error model.

Example
Explore diet quality for adolescents in the U.S. by looking at associations between added sugar and nutrient intakes.
Simple Linear Regression:

What if we let:

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

for \( i = 1, \ldots, n \) and \( \epsilon_i \sim N(0, \sigma_{ee}) \).

We estimate our parameters by minimizing the SSE:

\[ \text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \]

to get:

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]

\[ \hat{\beta}_0 = \bar{Y} - \bar{X} \hat{\beta}_1 \]
Motivation - An Illustration

But instead we actually observe: \( X_{ij} = x_i + u_{ij} \) where we have the measurement error \( u_{ij} \sim N(0, \sigma_{uu}) \).

When measurement error is present, if we estimate \( \beta_0 \) and \( \beta_1 \) using classical least squares estimation, these estimates will be biased towards zero.

Figure: Classical least squares fitted line (red) and MEM fit (blue).
Going to let $Y_{ij}$ be the observed nutrient intake for the $i$th individual on the $j$th day.

But we are interested in $y_i$, the \textbf{unobservable usual} (long term average) intake of a nutrient.

Let:

\[
Y_{ij} = y_i + w_{ij} \quad y_i \sim N(\mu_y, \sigma_{yy}) \quad w_{ij} \sim N(0, \sigma_{ww})
\]

where $i=1,...,n$ individuals and $j=1,2$ days per individual. The values $\sigma_{yy}$ and $\sigma_{ww}$ denote the between and within person variances and $y_i$ and $w_{ij}$ assumed independent.
Methodology — Measurement Error Model

Both the response and predictors have measurement error attached to the observed intakes.

\[ Y_{ij} = y_i + w_{ij} \]

\[ X_{ij} = x_i + u_{ij}, \]

where \( E(Y_{ij} | i) = y_i \) and \( E(X_{ij} | i) = x_i \).
Hence we get the model:

\[ Y_{ij} = \beta_0 + x_i \beta_1 + e_{ij}, \]

where \( e_{ij} = q_i + w_{ij} \) is composed of two independent parts:

- \( q_i \): The error in measuring \( y_i \) since observations vary from day to day
- \( w_{ij} \): The error in the equation when measuring usual intake

We also note the **possible correlation** between our observed response and predictor variables such that:

\[
\begin{bmatrix}
  w_{ij} \\
  u_{ij}
\end{bmatrix} \sim N \left( \begin{bmatrix}
  0 \\
  0
\end{bmatrix}, \begin{bmatrix}
  \sigma_{ww} & \sigma_{wu} \\
  \sigma_{wu} & \sigma_{uu}
\end{bmatrix} \right).
\]
Then the unbiased estimates of the averaged variances and covariances of the individuals follow as:

\[ \hat{\sigma}_{ww} = \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}_{ww,i} = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_i})^2}{n_i - 1} \right] = \frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{j=1}^{2} (Y_{ij} - \overline{Y_i})^2 \right] \]

where \( \hat{\sigma}_{uu} \) and \( \hat{\sigma}_{wu} \) are calculated similarly and \( n_i = 2 \) in our case.

Let the sample means for the predictors and response variables per each individual are defined as:

\[ \overline{X}_i. = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} = \frac{X_{i1} + X_{i2}}{2} \quad \overline{Y}_i. = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{Y_{i1} + Y_{i2}}{2}. \]
Methodology — Linear Regression Estimates

Regression coefficients are adjusted to account for:

- The error attached to our observed, measured daily intakes for both response and predictor variables
- Any correlation between response and predictor variables
Thus we calculate our regression coefficients as:

\[
\hat{\beta}_1 = \frac{M_{XY} - \hat{\sigma}_{wu}}{M_{XX} - \hat{\sigma}_{uu}}; \quad \hat{\beta}_0 = \bar{Y} - \bar{X} \cdot \hat{\beta}_1
\]

Where:

- \(M_{XX}\) and \(M_{XY}\) be the variance and covariance, respectively, of the averaged, usual intakes per individual such that:

\[
M_{XX} = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{X}_i - \bar{X})^2 \quad \text{and} \quad M_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{X}_i - \bar{X})(\bar{Y}_i - \bar{Y})
\]

- \(\hat{\sigma}_{wu}\) and \(\hat{\sigma}_{uu}\) are the averaged variance and covariances of intakes for individuals — A robust estimate is computed using medians to avoid poor model fit.

- \(\bar{Y}\) and \(\bar{X}\) are the overall averages of the response and predictor variables.
Methodology — A note of variance estimates

The strongly skewed distributions of $\hat{\sigma}_{wwi}$, $\hat{\sigma}_{uui}$ and $\hat{\sigma}_{wui}$:

Figure: Example of skewed distribution of $\hat{\sigma}$ values.

led us to use the median of the within individual variances and covariances for more robust estimates to avoid poor model fit.
Extensions

This methodology can also be extended to:

- More predictors — with or without measurement error

- Include survey weights
  - To produce data representative of the population

- Nonlinear case
The Data

Data from **NHANES (National Health and Nutrition Examination Surveys)**

- Uses a 24-hr recall method to collect daily food intake information on individuals
- We have 2 independent days of data per individual
- Food consumption then converted to nutrient intakes through an extensive table developed by U.S. Dept of Agriculture
- Empty calorie information (e.g. added sugar intakes) computed separately and recorded in MyPyramid Equivalents Database
Goal: To explore any associations that may exist between consumption of added sugars and nutrient intakes for children and adolescents ages 9-18y.

Previous Work: Conflicting results

- No clear negative association found yet
- Debates on how to control for total energy
**Empty Calories:** Calories that are contributed by some starchy foods, saturated fats, alcohol, and refined sugars. (Jenkins 2004)

**Added Sugars:** Any sugars eaten separately, used as an ingredient in processed or prepared foods, or added to foods after they are prepared.

**Discretionary (Solid) Fats:** Any fat in excess of the allowable amount that are solids at room temperature.
Exploratory Analysis of Empty Calories

**DCA:** (Discretionary Calorie Allowance) Any remaining calories after all nutrient-dense foods have been accounted for in each food group (e.g. an upper limit for the solid fats and added sugars recommended given an individual’s USDA Food Guide calorie level). *(J. Reedy & S. Krebs-Smith 2010)*

**USDA Food Guide calorie levels:**

<table>
<thead>
<tr>
<th>Calorie Level</th>
<th>1,000</th>
<th>1,400</th>
<th>1,800</th>
<th>2,200</th>
<th>2,600</th>
<th>3,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discretionary Calories</td>
<td>165</td>
<td>171</td>
<td>195</td>
<td>290</td>
<td>410</td>
<td>512</td>
</tr>
</tbody>
</table>

**Table:** Sample of the table with the number of calories that remain for each calorie level if nutrient dense foods are selected. Ideally, the total amount of discretionary calories should be limited to the amounts shown here for a given calorie level of an individual.

Using this table we come up with a predicted DCA for *any* calorie level.
Figure: Comparing linear (red), quadratic (green), and LOESS (blue) fits for regressing discretionary calories on the USDA Food Guide calorie level.

\[ \hat{\text{DiscCal}} = \hat{\beta}_0 + \hat{\beta}_1(\text{CalLevel}) + \hat{\beta}_2(\text{CalLevel})^2 \]

\[ = 299.3 - 0.2484x + 0.0001x^2 \]
Exploratory Analysis of Empty Calories

We can then predict the DCA for all individuals in our dataset and compare to their observed empty calorie intake:

Figure: Looking at the difference in the amount of actual discretionary calories consumed in comparison to an individual’s predicted DCA. Most (94.76%) individuals exceed their allowance.
Model Setup

Variables of interest:

Responses:
- Nutrient densities (e.g. Calcium (mg) per total caloric intake)
- Natural log transformations were taken for more symmetrical distributions.

Predictors:
- Added Sugar ($E_{AS}$) and Discretionary Solid Fat ($E_{SOL}$)
  - Accounting for total energy
- Age (years), BMI, weekend and race/ethnicity indicators
  - Consider separate age-sex groups: Males and Females 9-13y, Males and Females 14-18y
The Model

Recall:

- Only 2 days data per individual
- Interest is in the longterm average intakes
- Will use a measurement error model (Fuller 1987) to reduce bias in estimated regression coefficients attributed to nuisance day-to-day variation
- Survey weights incorporated into analysis
Results — Variance of Coefficients

In order to proceed with any inference about our coefficients we must consider the variance of our estimates.

Recall we must account for both:

- Measurement Error
- Survey Weights
Basic Jackknife Estimator:

- Resampling technique to estimate the precision of $\hat{\beta}$
- Leaving out one (or more) observations each time from the sample

$$\overline{\text{Var}}(\beta) = \frac{n}{n - 1} \sum_{i=1}^{n} (\hat{\beta}(-i) - \hat{\beta})^2$$

Advantage with survey data:

- Population split by strata (e.g. geographical region, minority populations, etc.)
- Sample from two primary sampling units (PSUs) within each strata
Thus, we use a fractional jackknife estimate:

\[ \hat{V}_{FJK}(\hat{\beta}_1) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=i}^{n_h} (\hat{\beta}_{(hj)} - \hat{\beta}_{\text{full}})(\hat{\beta}_{(hj)} - \hat{\beta}_{\text{full}}) \]

where:

- \( n_h \) is the number of PSUs in stratum \( h \) — two in our case.

- The estimate \( \hat{\beta}_{\text{full}} \) is computed with the original weights.

- The estimates \( \hat{\beta}_{(hj)} \) are computed using adjusted weights such that individuals in PSU \( j \) of stratum \( h \) are weighted lower than those in stratum \( h \) but not PSU \( j \); weights not in stratum \( h \) are left alone.
Results — Statistically Significant

Variances of our estimates now allow us to test whether or not our regression coefficient estimates are significant.

\[ H_0 : \beta = 0 \quad \text{\( H_A : \beta \neq 0 \)} \]

Calculate our test statistic as:

\[ t = \frac{\hat{\beta}}{SE_{\hat{\beta}}} \]

Where we let \( SE_{\hat{\beta}} = \sqrt{\hat{V}_{FJK}(\hat{\beta}_1)} \)

And come up with a p-value — hopefully less than \( \alpha = .05 \) so may reject \( H_0 \).
Results — Regression Coefficients Example

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.8196</td>
<td>0.0962</td>
<td>39.7184</td>
<td>0.0000</td>
</tr>
<tr>
<td><strong>Added Sugar</strong></td>
<td>−0.0051</td>
<td>0.0007</td>
<td>7.3444</td>
<td>0.0000**</td>
</tr>
<tr>
<td>Solid Fat</td>
<td>0.0041</td>
<td>0.0006</td>
<td>7.2503</td>
<td>0.0000**</td>
</tr>
<tr>
<td>BMI</td>
<td>0.0033</td>
<td>0.0014</td>
<td>2.3428</td>
<td>0.0192</td>
</tr>
<tr>
<td>Age</td>
<td>−0.0079</td>
<td>0.0055</td>
<td>1.4202</td>
<td>0.1557</td>
</tr>
<tr>
<td>Weekend</td>
<td>−0.0507</td>
<td>0.0328</td>
<td>1.5439</td>
<td>0.1228</td>
</tr>
<tr>
<td>Black</td>
<td>−0.2409</td>
<td>0.0233</td>
<td>10.3455</td>
<td>0.0000</td>
</tr>
<tr>
<td>MexAm</td>
<td>−0.0466</td>
<td>0.0278</td>
<td>1.6775</td>
<td>0.0936</td>
</tr>
<tr>
<td>Hisp</td>
<td>−0.1000</td>
<td>0.0379</td>
<td>2.6369</td>
<td>0.0084</td>
</tr>
<tr>
<td>Other</td>
<td>−0.0812</td>
<td>0.0420</td>
<td>1.9335</td>
<td>0.0533</td>
</tr>
</tbody>
</table>

Table: Regression estimates for Female 14-18y and Calcium.
Results — Correlations

### Added Sugar Correlations

<table>
<thead>
<tr>
<th>Nutrient</th>
<th>Male 9-13y</th>
<th>Male 14-18y</th>
<th>Female 9-13y</th>
<th>Female 14-18y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium</td>
<td>−0.096</td>
<td>−0.144</td>
<td>−0.086</td>
<td>−0.123</td>
</tr>
<tr>
<td>Zinc</td>
<td>−0.083</td>
<td>−0.069</td>
<td>0.003</td>
<td>−0.059</td>
</tr>
<tr>
<td>Folate</td>
<td>−0.044</td>
<td>0.073</td>
<td>−0.073</td>
<td>−0.023</td>
</tr>
<tr>
<td>Fiber</td>
<td>−0.078</td>
<td>−0.034</td>
<td>−0.070</td>
<td>−0.053</td>
</tr>
<tr>
<td>Vitamin E</td>
<td>−0.076</td>
<td>−0.099</td>
<td>−0.081</td>
<td>−0.107</td>
</tr>
<tr>
<td>Magnesium</td>
<td>−0.160</td>
<td>−0.126</td>
<td>−0.042</td>
<td>−0.063</td>
</tr>
<tr>
<td>Phosphorus</td>
<td>−0.261</td>
<td>−0.245</td>
<td>−0.131</td>
<td>−0.246</td>
</tr>
<tr>
<td>Sodium</td>
<td>−0.303</td>
<td>−0.274</td>
<td>−0.198</td>
<td>−0.247</td>
</tr>
<tr>
<td>Potassium</td>
<td>−0.222</td>
<td>−0.133</td>
<td>−0.134</td>
<td>−0.086</td>
</tr>
</tbody>
</table>

**Table:** Added sugar correlations after accounting for survey weights as well as measurement error. Correlations with corresponding significant parameter estimates (at $\alpha = .05$ level) are shaded.
Discussion

- We used a model that accounted for:
  - Total Energy
  - Measurement Error
  - Survey Weights

- Found results supporting a negative association between added sugar and intakes of certain essential nutrients for individuals ages 9-18y in the U.S.
Extensions — Another Example

Motivation:
- Low levels of vitamin D are associated with poor bone health.
  - To reduce risks, want to improve food and supplement intake guidelines.

- Vitamin D levels influenced by more than food and supplement intake. (e.g. sun exposure)
  - Serum 25-hydroxy vitamin D (OHD) is a biomarker that monitors levels of vitamin D.

Goal:
- Ideally want vitamin D recommendations for food and supplement intake (not for levels of a biomarker).

- Investigate the association between levels of OHD with levels of vitamin D intake.
Another Example — The Data

- Data collected on 3,868 individuals
- 1-2 days of vitamin D (food and supplement) intake data per individual
- One value of OHD per individual
- Three ethnicity groups: whites, blacks, and other
Another Example — Exploratory

- To achieve more symmetric distributions we will transform our data.
  - $y_i = \text{OHD}^{1/2}$ and $w_{ij} = \text{VitD}^{1/4}$ for $i = 1, \ldots, n$ individuals and $j = 1, 2$ days

- Preliminary Loess curves suggested an 'S' shaped curve.
  - Suggests a mean function of the form:
    \[
    f(x_i, \beta) = K + \frac{B}{1 + \exp(a(x_i - m))}
    \]
Again, we are ideally interested in the unobservable, usual intake of vitamin D and should account for nuisance variation from estimating usual intake with just two days of data.

We now consider our predictor to be measured with error:

\[ w_{ij} = x_i + u_{ij} \]

for

\[ u_{ij} \sim N(0, \sigma_{uu}) \]

where we assume \( \sigma_{uu} \) to be fixed as the average of individual variances between days.
Another Example — Model with error in the predictor

- We must also assume a distribution for the unknown, usual vitamin D intake:
  \[ x_i \sim N(\mu_x, \sigma_{xx}). \]

The model for OHD levels then depends on these usual intakes of vitamin D:

\[
y_i = K + \frac{B}{1 + \exp(a(x_i - m))} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_{ee})
\]
Another Example — Likelihood

Parameter estimates found by maximizing the likelihood. Where in general we have:

\[
f(y_i, w_{ij} | \beta, x_i) = \int_x f(y_i | x_i, \beta)f(w_{ij} | x_i)f(x)dx
\]

Giving us the likelihood (assuming all Normal distributions):

\[
L_i(\theta) = \frac{1}{\sqrt{\sigma_{uu} \sigma_{ee} \sigma_{xx}}} \int_x \phi \left( \frac{y_i - f(x_i, \beta)}{\sigma_{ee}^{1/2}} \right) \phi \left( \frac{w_{ij} - x_i}{\sigma_{uu}^{1/2}} \right) \phi \left( \frac{x_i - \mu_x}{\sigma_{xx}^{1/2}} \right) dx
\]

Which may be approximated by numerical integration:

\[
L_i(\theta) \approx \frac{1}{M} \sum_{k=1}^M \frac{1}{\sqrt{\sigma_{uu} \sigma_{ee} \sigma_{xx}}} \phi_k \left( \frac{y_i - f(x_{ki}, \beta)}{\sigma_{ee}^{1/2}} \right) \phi_k \left( \frac{\bar{w}_{i.} - x_{ki}}{\sigma_{uu}^{1/2} / 2} \right) \phi_k \left( \frac{x_{ki} - \mu_{x_k}}{\sigma_{xx_k}^{1/2}} \right)
\]
Another Example — Results

- K and B estimated separately by ethnicity

Figure: Nonlinear measurement error model fit.
Calcium

- Another nutrient related to bone health
- Preliminary analysis shows calcium intakes highly correlated with vitamin D intakes

Next Step: Add more predictors such as calcium intake to our nonlinear model.
Accredgments:

Special thanks to Alicia Carriquiry (ISU) and Wayne Fuller (ISU).
Also thanks to Ulrike Genschel (ISU), Helen Jensen (ISU), and Patricia Guenther (CDC) for their input.