ESTIMATION OF DAY-TO-DAY CORRELATIONS FOR DIETARY COMPONENTS

Alicia L. Carriquiry and Wayne A. Fuller, Iowa State University
Alicia L. Carriquiry, Iowa State University, Ames, IA 50011

KEY WORDS: Nutrition, ratio estimates, factor analysis, usual intakes.

1 INTRODUCTION

The Continuing Survey of Food Intake by Individuals, (CSFII) is a nationwide food consumption survey conducted by the Agricultural Research Service of the U.S. Department of Agriculture. In this survey, individuals are contacted and asked to report their food intakes on three consecutive days. The reported food intake data are converted to intakes of dietary components such as protein, fat, and vitamin A, using food conversion databases.

We are interested in the distribution of usual intakes, where the individual's usual intake is defined to be the long run average intake of a nutrient for that particular individual. Also of interest is the fraction of individuals whose intake of a dietary component such as protein is above or below a particular point (e.g., the recommended daily allowance, RDA). Nusser et al. (1995) present a method for estimating the distribution of usual nutrient intake. The estimation of the distribution of usual intakes requires the estimation of the day-to-day variance of individual intakes and also of the individual-to-individual variance in intakes. To estimate these variance components, the correlation among the three-day intakes for a particular individual needs to be accounted for.

In this analysis, we will use data on approximately 12,000 individuals divided into five sex-age groups to obtain an estimator of the day-to-day correlation in intakes for several dietary components. The five sex-age groups are males twenty to fifty-nine, females twenty to fifty-nine, males sixty and over, females sixty and over, and individuals nineteen years of age and younger. We will consider thirty-four dietary components in the analysis. Twenty-seven of the components are nutrients such as carbohydrates, calcium, and protein. In addition, water, and six ratios of dietary components are analyzed. The six ratios include the percent of total calories consumed that are attributable to protein, carbohydrates, and various types of fat.

2 THE MODEL

The basic model is a Gaussian measurement error model that postulates that the observed intake on the j-th day for the i-th individual is the sum of a mean, an individual effect, and a measurement error. Thus, the model is

$$\begin{aligned} X_{ij} &= \mu + \gamma_i + \epsilon_{ij}, \qquad (1) \\ \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix} &\sim NI\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \sigma_\epsilon^2 \right), \end{aligned}$$

where X_{ij} is the (power transformed) intake for the i-th individual on the j-th day, $j = 1, 2, 3$, and γ_i is the individual effect. The data are power transformed so that they are approximately normally distributed. Ultimately, we are interested in the distribution of the usual intakes, $\mu + \gamma_i$. Note that under model (1), the covariance matrix of the daily deviations is the covariance function of a first order autoregressive model.

The model states that the observed intake X_{ij}, conditional on the i-th individual, is unbiased for that individual's usual intake. It is well known that if X_{ij} represents an observed ratio, say X_{1ij}/X_{2ij}, then the unbiasedness assumption is not satisfied. Therefore, we obtain estimators for the ratios such that the estimated ratios fit into the same analysis structure as individual dietary components. We first estimate the ratio for an individual, and then obtain three "daily" values for each individual in the sample. Letting $\hat{R}_i$ denote the estimated ratio for the i-th individual, we have

$$\hat{R}_i = \frac{\bar{x}_{1i}\bar{x}_{2i}^{-1} + \hat{C}\{\bar{x}_{1i}, \bar{x}_{2i}\}}{\bar{x}_{2i}^2 + \hat{V}\{\bar{x}_{2i}\}},$$

where
$(\bar{x}_{1i}, \bar{x}_{2i}) = 3^{-1}\sum_{j=1}^{3}(X_{1ij}, X_{2ij})$, $\hat{C}\{\bar{x}_{1i}, \bar{x}_{2i}\}$ is

the estimated covariance between $\bar{X}_{1i}$ and $\bar{X}_{2i}$, and $\hat{V}\{\bar{x}_{2i}\}$ is the estimated variance of $\bar{x}_{2i}$. In the ratio, X_1 is the nutrient in the numerator. The estimated ratio is Beal's estimator (Beale (1963)) which is known to be less biased than the simple estimator $\bar{x}/\bar{x}_2^{-1}$ (e.g., Cochran, 1977). We obtain three individual observations from $\hat{R}_i$ as

$$\tilde{R}_{ij} = \hat{R}_i + \left(\bar{x}_{2i} + \hat{V}\{\bar{x}_{2i}\}\right)^{-1} \left(X_{1ij} - \bar{x}_{2i}^{-1}\bar{x}_{1i}X_{2ij}\right), \quad (2)$$

where the quantity added to $\hat{R}_i$ is the Taylor approximation to the error for an individual on a particular day. Under this approach, the mean of the three observations for an individual is equal to Beal's estimator and the estimated variance of the three observations is the usual Taylor approximation to the variance of the ratio. Details on this procedure are given by Carriquiry et al. (1995).

3 Direct Estimation of ρ

We now present a procedure for estimating the day-to-day correlation coefficient ρ, from observed intakes X_{ij}, or from estimated ratios $\tilde{R}_{ij}$. This method was given by Carriquiry et al. (1995).

To estimate the correlations, we consider two linear combinations of the three observations for each individual,

$$\begin{aligned} L_{1i} &= X_{i1} - X_{i3}, \\ L_{2i} &= X_{i1} - 2X_{i2} + X_{i3}. \end{aligned}$$

Under the model,

$$\mathrm{Var}(L_1)\left[\mathrm{Var}(L_2)\right]^{-1} = (1+\rho)(3-\rho)^{-1}. \quad (3)$$

It follows that an estimator of ρ can be constructed from (3) by replacing the unknown variances $\mathrm{Var}(L_1)$, $\mathrm{Var}(L_2)$ by the corresponding estimated variances. An estimator of ρ is then given by

$$\hat{\rho} = \frac{3\widehat{\mathrm{Var}}(L_1) - \widehat{\mathrm{Var}}(L_2)}{\widehat{\mathrm{Var}}(L_1) + \widehat{\mathrm{Var}}(L_2)},$$

where $\widehat{\mathrm{Var}}(L_i)$ is the usual sample variance of L_1.

Under the assumption that $\hat{\rho}$ is approximately normally distributed, the Taylor series estimator of the variance of $\hat{\rho}$ is

$$\hat{V}\{\hat{\rho}\} = [4(n-1)]^{-1}(1+\hat{\rho})^2(3-\hat{\rho})^2.$$

We applied this procedure to intake data obtained during 1989, 1990, and 1991 in the CSFII. Intake data were divided into age-sex groups as discussed earlier, and direct correlation estimates $\hat{\rho}$ were obtained for each of 34 dietary components in each sex-age group.

Table 1 contains the average $\hat{\rho}$'s (over 34 dietary components) for the five sex-age groups as well as the average standard errors of $\hat{\rho}$. These are a function of the number of observations in a group and, to a modest extent, of the actual estimated correlations. The third line of the table contains the standard deviations of the $\hat{\rho}$ across the components for a particular group. Notice that for males, the standard deviation is not much larger than the estimation error in $\hat{\rho}$. For females and younger persons, the standard deviation is about twice the estimation error in $\hat{\rho}$. The estimated correlations are moderately small and there is a tendency for the estimates associated with a particular dietary component to be of similar magnitude.

4 Smoothed Estimates of ρ

Results obtained in the preceding section (see, e.g., Table 1) suggest a model in which the correlations in different subpopulations are similar and are linear combinations of some underlying unobservable, common correlation coefficients.

After some experimentation, we chose a two-factor model to represent the direct correlation estimates

$$\begin{aligned} \rho_{k\ell} - \mu_\ell &= \beta_{i\ell}(\rho_{k4} - \mu_4) + \beta_{2\ell}(\rho_{k5} - \mu_5), \\ \hat{\rho}_{k\ell} &= \rho_{k\ell} + e_{k\ell} \end{aligned} \quad (4)$$

for $k = 1, 2, ..., 3, 4$ and $\ell = 1, 2, ..., 5$, where k denotes dietary components and ℓ denotes subpopulations. See, e.g., Fuller (1987). In model (4), $\hat{\rho}_{k\ell}$ denotes the direct correlation estimate for the k-th component in the ℓ-th sex-age group. In the model, the correlations for the first three subpopulations are written as a linear function of the correlations in the second two subpopulations. This choice is arbitrary, and produces the same result as any other choice. The errors in the direct correlation estimates, denoted by $e_{k\ell}$, are the estimation errors.

The estimated values for the coefficients in model

Table 1. Average estimated correlations and standard errors.

	Males 20-59	Males ≥ 60	Females 20-59	Females ≥ 60	Persons 0-19
Ave($\hat{\rho}$)	0.098	0.138	0.097	0.124	0.069
Ave(se$\hat{\rho}$)	(0.032)	(0.055)	(0.028)	(0.042)	(0.026)
s.d.($\hat{\rho}$)	(0.035)	(0.066)	(0.046)	(0.093)	(0.051)

Table 2. Direct and model estimates for 34 components ($\times 100$)

Group	Direct estimate	Smooth estimate	Reduction
Males 20-59	3.2	0.6	80%
Males ≥ 60	5.5	2.0	63%
Females 20-59	2.8	1.4	51%
Females ≥ 60	4.2	3.8	10%
Persons 0-19	2.6	2.2	13%

(4), and their standard errors, are

$$\begin{pmatrix} \hat{\beta}_{11} & \hat{\beta}_{12} & \hat{\beta}_{13} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \hat{\beta}_{23} \end{pmatrix} = \begin{pmatrix} 0.49 & 0.50 & 0.17 \\ (0.09) & (0.16) & (0.09) \\ 0.28 & -0.17 & 0.58 \\ (0.17) & (0.31) & (0.17) \end{pmatrix}.$$

The F-test for the adequacy of the 2-factor model is $F = 1.16$ with 93 and infinity degrees of freedom. On the basis of the F-test, we conclude that the 2-factor model gives an adequate representation of the data. Further, smoothed correlation estimates obtained from the factor model are much less variable than the direct estimates. Table 2 contains estimated standard errors for the direct and the model estimates. The largest gains in precision are achieved in the two male groups. There is considerable gain for females aged twenty to fifty-nine, and only modest gains in efficiency for females older than fifty-nine, and persons nineteen and younger. These latter two categories are the categories where the component-to-component variation was fairly large relative to the estimation error.

5 Conclusions

From a subject matter viewpoint, the fact that day-to-day correlations for most dietary components in all age-sex groups are small in absolute magnitude has considerable importance. This suggests, for example, that collecting data on consecutive days is only slightly less efficient than collecting it on independent days.

Using the factor model to obtain smoothed estimates of day-to-day correlations for dietary components, including ratios of dietary components, gave estimates that are more efficient than the direct estimates.

While differences between the day-to-day correlations for the various dietary components and various age-sex groups are statistically significant, they are small in absolute value.

ACKNOWLEDGMENT

This research was partly supported by Cooperative Agreement 58-3198-2-006 with the Agricultural Research Service, U.S. Department of Agriculture.

References

Beale, E. M. L. (1962). Some uses of computers in operational research. *Industrielle Organization*, 31, 27-28.

Carriquiry, A. L., Fuller, W. A., Goyeneche, J. J., and Jensen, H. H. (1995). Estimated correlations among days for the combined 1989-91 CSFII. *Research report prepared for the Agricultural Research Service, U.S. Dept. of Agriculture. Department of Statistics, Iowa State University.* June, 1995.

Carriquiry, A. L., Fuller, W. A., Goneyeche, J. J., and Dodd, K. W. (1995). Estimation of the usual intake distributions of ratios of dietary components. *Research report prepared for the Agricultural Research Service, U.S. Dept. of Agriculture. Department of Statistics, Iowa State University.* September, 1995.

Cochran, W. G. (1977). *Sampling Techniques.* 3rd ed. Wiley.

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

Nusser, S. M., Carriquiry, A. L., Fuller, W. A., and Dodd, K. W. (1995). A semiparametric