Estimation in the Presence of Measurement Error

Dietary Assessment Research Series 1

Wayne A. Fuller

Staff Report 94-SR 73
December 1994

Center for Agricultural and Rural Development
Iowa State University
Ames, Iowa 50011

Wayne A. Fuller is distinguished professor, Department of Statistics, Iowa State University.

This research was partly supported by Research Support Agreement 58-3198-9-032 and Cooperative Agreement 58-3198-2-006 with the Human Nutrition Information Service, U.S. Department of Agriculture and Joint Statistical Agreement JSA91-21 with the U.S. Bureau of the Census.

CONTENTS

FIGURES

TABLES

FOREWORD

The Dietary Assessment Research Series presents research and methods developed to better evaluate survey data used to assess the quality of diets. Increased scientific evidence links dietary intakes to health outcomes. Recent concerns about exposure to foodborne contaminants is also motivating increased attention to improved methods of dietary assessment. Improved capacities for dietary assessment can form a stronger foundation for considering policies on food safety, food technology, public health, labor productivity, and other areas related to human health and performance.

Much of the research published in this series is collaborative among researchers in CARD and the Department of Statistics, and nutrition scientists at Iowa State University and staff at the U.S. Department of Agriculture, Human Nutrition Information Service (now Agricultural Research Service). The identification of the series will make it possible for those interested in the results of the research to more easily access them. Also, we hope that it will attract broader professional attention to the important area for scientific and policy research.

The work at CARD and the Department of Statistics is designed to improve statistical, survey, and research methods used in dietary assessments, and to foster better understanding of diet, nutrition, and health consequences of food consumption and eating patterns. This series reports developments in methods, including documentation and applications, that will contribute to better understanding of dietary intakes, exposures to foodborne contaminants, the quality of diets, and factors associated with the outcomes. The series also will contain reports related to food and health policy issues that can be better addressed with new technology.

ABSTRACT

The importance of measurement error for parameter estimation and for the design of statistical studies, particularly sample surveys, is examined. Beginning with a brief review of Hansen's contributions, the discussion concentrates on estimation problems in which measurement error leads to bias in the usual estimators. Estimation of distribution functions and regression equations are discussed, and the implications for the design of surveys are presented.

ESTIMATION IN THE PRESENCE OF MEASUREMENT ERROR

1. Introduction

That data available for statistical analysis are subject to error is universally recognized. Two recent books, Nonsampling Error in Surveys by Lessler and Kalsbeek (1992) and Measurement Errors in Surveys edited by Biemer, Groves, Lyberg, Mathiowetz, and Sudman (1991), are evidence of the importance attached to measurement error by the survey sampling community. The papers on this topic at the August 1992 meeting of the American Statistical Association also demonstrate the concern for the measurement process in survey sampling.

The majority of the works cited are devoted to identifying sources of error, developing models for measurement error, designing procedures to minimize measurement error (or to minimize total error), and to measuring the properties of measurement error. The survey sampling literature contains less material on the analyses of data observed subject to measurement error. In other disciplines where survey data are used, notably psychology, sociology, and epidemiology, measurement variability is an integral part of the analysis. We shall address some aspects of the analysis problem.

2. Models for Response Error

Hansen and his co–workers at the U.S. Bureau of the Census conducted extensive studies of response error in the 1950's and 1960's. They assumed that the observations in a survey are the result of a trial from the set of all possible trials. They also assumed that there is a true value for the characteristic of interest for each individual. Let $\hat{\theta}$ be an estimator constructed from a survey and let θ^0 be the true value. Let $E\{\hat{\theta}|s\}$ be the

expected value of the estimator over all possible trials for a particular sample, s. Hansen, Hurwitz, and Bershad (1961) suggested the decomposition,

$$\begin{aligned} E\{(\hat{\theta} - \theta^0)^2\} &= [E\{(\hat{\theta} - \theta^0)\}]^2 \\ &+ E\{[\hat{\theta} - E(\hat{\theta}|s)]^2\} \\ &+ E\{[E(\hat{\theta}|s) - E(\hat{\theta})]^2\} , \end{aligned} \tag{2.1}$$

where the first term on the right of the equality is the squared bias in the estimator, the second term is the response variance (measurement variance) contribution, and the final term is the sampling variance of the estimator. In their introduction the authors also included the covariance between the response variance and the sampling variance, but the covariance term was not included in the remainder of their discussion. To investigate the response variance, Hansen, Hurwitz, and Bershad (1961) defined the response of the j–th unit at trial t of sample s by

$$Y_{jts} = y_j + w_{jts} , \tag{2.2}$$

where $y_j = E\{y_{jts}\}$ is the expected value of the response of individual j over all possible trials of samples containing unit j. The response error w_{jts}, called the response deviation, by Hansen, Hurwitz, and Bershad (1961), has zero expected value by definition, but the expected value for a particular sample, $E\{w_{jts}|s = s^*\}$ need not be zero. Hansen, Hurwitz, and Bershad (1961) were particularly interested in the correlation between the w_{jts} for different individuals in the same sample. In interviewer surveys, there is generally a positive correlation among respondents interviewed by the same individual. Hansen, Hurwitz, and Bershad (1961) give some estimates of the correlation obtained from the

response variance study conducted in conjunction with the 1950 U.S. Census of Population and Housing.

Hansen and his co–workers were concerned with censuses and large surveys and with relatively simple estimators, such as estimators of means and totals. The emphasis in this early work was on efficient survey design. It exemplifies research at its finest. Areas of importance were identified, data were collected, and decisions made on the basis of the analysis of the data.

There is today a great deal of research on measurement errors in surveys. For survey statisticians, the emphasis remains much as it was during Hansen's early work. Attempts are made to identify and quantify sources of response variability and response bias, and to use this information in creating efficient survey designs for the estimation of totals and means. Survey statisticians are concerned about response error at the design stage, but response error is less often considered at the estimation stage. Also, it is not a part of survey convention to collect information on response error for the important items in a survey. Rather special studies may be conducted, often before the actual survey.

A reason for the present custom is the emphasis given to means and totals in survey methodology. Assume: (i) the response error is unbiased, (ii) response errors in different primary sampling units are uncorrelated, and (iii), the finite correction term can be ignored. Then, the usual survey estimators of the mean and of the variance of the mean are unbiased. These results are stated by Cochran (1977, p. 396).

> "Errors of measurement that are independent from unit to unit within the sample and average to zero over the whole population are properly taken into account in the usual formulas for computing the standard errors of the estimates, provided that fpc terms are negligible. Such errors decrease the precision of the estimates, and it is worthwhile to find out whether this decrease is serious.
>
> If errors of measurement on different units in the sample are correlated, the usual formulas for the standard errors are biased. The standard errors are likely to be too small, since the correlations are mostly positive in practice. This type of disturbance is easily overlooked and may often have passed unnoticed."

We shall adopt a simplified model for measurement error and shall discuss the effect of measurement error on relatively common estimation procedures. Let the p–dimensional row vector of observations be denoted by

$$\mathbf{Z}_t = \mathbf{z}_t + \epsilon_t \,, \tag{2.3}$$

where $\mathbf{z}_t$ is the true value for unit t, and ϵ_t is the vector of measurement errors. We begin by assuming

$$\mathrm{E}\{\epsilon_t\} = 0 \text{ and } \mathrm{E}\{\epsilon_t \epsilon_t'\} = \Sigma_{\epsilon\epsilon} \tag{2.4}$$

for all t, and assume $\mathbf{z}_t$ is a vector random variable with mean

$$\mu_z = \mathrm{E}\{\mathbf{z}_t\} \tag{2.5}$$

and covariance matrix

$$\Sigma_{zz} = \mathrm{E}\{(\mathbf{z}_t - \mu_z)'(\mathbf{z}_t - \mu_z)\} \,. \tag{2.6}$$

Then $\mathbf{Z}_t$ is a vector random variable with mean μ_z and covariance matrix

$$\Sigma_{ZZ} = \mathrm{E}\{(\mathbf{Z}_t - \mu_z)'(\mathbf{Z}_t - \mu_z)\} = \Sigma_{zz} + \Sigma_{\epsilon\epsilon} \,. \tag{2.7}$$

If $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is a random sample, then

$$\mathrm{E}\{\bar{\mathbf{Z}}\} = \mu_z, \tag{2.8}$$

$$E\{\hat{\Sigma}_{ZZ}\} = \Sigma_{ZZ} = \Sigma_{zz} + \Sigma_{\epsilon\epsilon}\,, \tag{2.9}$$

where

$$\bar{Z} = n^{-1} \sum_{t=1}^{n} Z_t\,,$$

$$\hat{\Sigma}_{ZZ} = (n-1)^{-1} \sum_{t=1}^{n} (Z_t - \bar{Z})'(Z_t - Z)\,.$$

Thus, consistent with the quote from Cochran, $E\{\bar{Z}\} = \mu_z$ and

$$\hat{V}\{\hat{\mu}_z\} = n^{-1}\hat{\Sigma}_{ZZ} \tag{2.10}$$

is unbiased for the variance of $\hat{\mu}_z$, $V\{\hat{\mu}_z\}$. We shall see that the mean is, essentially, the only statistic for which this result holds.

We shall restrict our discussion to zero mean measurement error. The bias associated with measurement error whose mean is not zero is clearly important. There are a number of techniques in the literature for adjusting for measurement error that does not have zero mean. The most common is two–phase estimation in which "true" values are obtained for a subsample of the original sample.

3. Estimation of the Distribution Function

The estimated mean of z is one of the few statistics that remains unbiased in the presence of nontrivial response error. Assume that z_{t1} and ϵ_{t1} are independent normal scalar random variables. Then $Z_{t1} \sim NI(\mu_z, \sigma_{zz11} + \sigma_{\epsilon\epsilon 11})$, where σ_{zz11} is the variance of z_1 and $\sigma_{\epsilon\epsilon 11}$ is the variance of ϵ_1. Because the normal distribution is completely determined by the mean and variance, and because the variance of Z is greater than the

variance of z, the cumulative distribution function of Z agrees with the cumulative distribution function of z only at the point μ_z. For the normal distribution, the quantiles of the distribution of $Z_t - \mu_z$ are multiples of the quantiles of $z_t - \mu_z$. That is,

$$Q_{Z-\mu_z}(a) = [\sigma^{-1}_{zz11}\sigma_{ZZ11}]^{1/2} Q_{z-\mu_z}(a) ,$$

where $Q_X(d)$ is the quantile function of X evaluated at d, the quantile function is the inverse of the cumulative distribution function, and we use both σ_z^2 and σ_{zz} for the variance of the random variable z. Thus, the sample cumulative distribution function of Z is a biased estimator of the cumulative distribution function of z except at μ_z.

The effect of measurement error on the cumulative distribution function is described in Table 1 and Table 2. The importance of the effect of measurement error depends upon the part of the distribution function that is of interest. As mentioned, there is no bias in the estimated median of the normal distribution. The relative error increases as one moves away from the mean. For example, a measurement error with variance equal to 15% of the

Table 1. Percentiles in the presence of measurement error. Normal distribution $\sigma_z^2 = 1$.

Probability	σ_ϵ^2						
(%)	0.00	0.05	0.10	0.15	0.25	1.00	2.00
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000
75	0.674	0.691	0.707	0.723	0.754	0.954	1.168
90	1.282	1.313	1.344	1.374	1.433	1.812	2.220
95	1.645	1.685	1.725	1.764	1.839	2.326	2.849
99	2.326	2.384	2.440	2.495	2.601	3.290	4.029

Table 2. Cumulative distribution function in the presence of measurement error. Normal distribution $\sigma_z^2 = 1$.

Z–value	σ_ϵ^2						
	0.00	0.05	0.10	0.15	0.25	1.00	2.00
0.000	0.500	0.500	0.500	0.500	0.500	0.500	0.500
0.675	0.750	0.745	0.740	0.735	0.727	0.683	0.652
1.282	0.900	0.895	0.889	0.884	0.874	0.818	0.770
1.645	0.950	0.946	0.942	0.937	0.929	0.878	0.829
2.327	0.990	0.988	0.987	0.985	0.981	0.950	0.910

true variance changes the fraction of the observed values greater than 1.645 from 5% to 6.3%. Thus, a 15% error variance produces a 25% change in this particular parameter.

This discussion illustrates that, under the simple measurement error model, some estimators that might be considered to be "means" are biased. Let (z_{t1}, ϵ_{t1}) be the normal independent scalar random variables and let $Z_{t1} = z_{t1} + \epsilon_{t1}$. Consider the random variables

$$Z_{t2} = 1 \text{ if } Z_{t1} < A_2$$

$$= 0 \text{ otherwise}$$

and

$$z_{t2} = 1 \text{ if } z_{t1} < A_2$$

$$= 0 \text{ otherwise .}$$

Assume that we wish to estimate the mean of z_{t2},

$$\mu_{z2} = P\{z_{t1} < A_2\}$$

for some A_2 not equal to μ_{z1}. Then $\bar{Z}_2$ is a biased estimator of the mean of z_2, where the bias is illustrated in Table 2. The bias results from the fact that the mean of $Z_{t2} - z_{t2}$ is not zero.

To estimate the distribution function of the underlying true normal variables, we assume that we have a sample of m individuals and that replicate observations are made on some individuals. Let

$$Z_{ij} = z_i + \epsilon_{ij}, \tag{3.1}$$

$$\begin{bmatrix} z_i \\ \epsilon_{ij} \end{bmatrix} \sim \text{NI}\left[\begin{bmatrix} \mu_z \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{zz} & 0 \\ 0 & \sigma_{\epsilon\epsilon} \end{bmatrix}\right].$$

We use the analysis of variance estimators to estimate the components of variance. The analysis of variance is given in Table 3. These estimators are not the most efficient, but they are easy to construct.

Table 3. Analysis of variance for random components model.

Source	df	SS	MS	EMS
Individuals	$d_B = m - 1$	$\sum_{i=1}^{m} k_i(\bar{Z}_{i.} - \bar{Z}_{..})^2$	B	$\sigma_\epsilon^2 + k_0\sigma_z^2$
Within	$d_W = \sum_{i=1}^{m} (k_i - 1)$	$\sum_{i=1}^{m}\sum_{j=1}^{k_i} (Z_{ij} - \bar{Z}_{i.})^2$	W	σ_ϵ^2

The estimators associated with the table are

$$\hat{\sigma}_\epsilon^2 = W = d_W^{-1} \sum_{i=1}^{m} \sum_{j=1}^{k_i} (Z_{ij} - \bar{Z}_{i.})^2$$

$$\hat{\mu}_z = \bar{Z}_{..} = m^{-1} \sum_{i=1}^{m} \bar{Z}_{i.} , \tag{3.2}$$

and

$$\hat{\sigma}_z^2 = k_0^{-1}(B - W) ,$$

where

$$\bar{Z}_{i.} = k_i^{-1} \sum_{j=1}^{k_i} Z_{ij} ,$$

$$k_0 = (m-1)^{-1}\left[n - n^{-1} \sum_{i=1}^{m} k_i^2\right]$$

and B denotes the between individual mean square. Under the assumption that z_i and ϵ_{ij} are normally and independently distributed,

$$V\{\hat{\mu}\} = m^{-1}\sigma_z^2 + m^{-2} \sum_{i=1}^{m} k_i^{-1}\sigma_\epsilon^2 , \tag{3.3}$$

$$V\{\hat{\sigma}_\epsilon^2\} = 2d_W^{-1}\sigma_\epsilon^4 , \tag{3.4}$$

and

$$V\{\hat{\sigma}_z^2\} \doteq 2k_0^{-2}\left[(m-1)^{-1}m^{-1} \sum_{i=1}^{m} (\sigma_z^2 + k_i^{-1}\sigma_\epsilon^2)^2 k_i^2 + d_W^{-1}\sigma_\epsilon^4\right] . \tag{3.5}$$

The estimated π–quantile for the distribution of z is

$$\hat{Q}(\pi) = \hat{\mu}_z + \hat{\sigma}_z \Phi^{-1}(\pi) , \tag{3.6}$$

where Φ is the standard normal cumulative distribution, and π is the probability defining the quantile. Using Taylor series arguments, an estimator of the variance of $\hat{Q}(\pi)$ is

$$\hat{V}\{\hat{Q}(\pi)\} = [\Phi^{-1}(\pi)]^2 \hat{V}\{\hat{\sigma}_z\} + \hat{V}\{\hat{\mu}_z\} , \tag{3.7}$$

where $\hat{V}\{\hat{\sigma}_z\} = (4\hat{\sigma}_x^2)^{-1}\hat{V}\{\hat{\sigma}_z^2\}$, and $\hat{V}\{\hat{\sigma}_z^2\}$ and $\hat{V}\{\hat{\mu}_z\}$ are obtained from (3.5) and (3.3) by replacing parameters with their estimators.

In some studies, a portion of the measurement error is due to sampling of the basic material. To illustrate this and to illustrate the estimation of the cumulative distribution function in the presence of measurement error, we use some data from the Human Nutrition Information Service of the U.S. Department of Agriculture. This research is described more fully in Nusser, Carriquiry, Dodd and Fuller (1994).

The data are a subset of the data from the 1985 Continuing Survey of Food Intakes by Individuals conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture. Daily dietary intakes were collected from women between 19 and 50 years of age. Daily intakes were obtained at approximate two–month intervals over the period, April 1985 to March 1986. Data for the first day were collected by personal interview and were based on a 24–hour recall. Data for subsequent days were based on 24–hour recall and were collected by telephone whenever possible. The sample was a multi–stage stratified area probability sample from the 48 coterminous states. The primary sampling units were area segments, and the probabilities of selection of area segments were proportional to the numbers of housing units in the segments as estimated by the Bureau of the Census. The sample was designed to be self weighting. Because of

the high rate of attrition over the six waves of data collection, the Human Nutrition Information Service constructed a four–day data set for analysis. The four days of data consisted of the first day of dietary intake for all individuals who provided at least four days of data, plus a random selection of three daily intakes from the remaining three, four or five days of data available. Weights were developed to adjust for nonresponse, but the analyses of this paper are constructed on unweighted data.

We analyze a subset of the four–day data set containing dietary intakes for women between 25 and 50 years of age who were responsible for meal planning or preparation within the household and who were not pregnant or lactating during the survey period. There were 737 women who belonged to this category. Because of the time separation of the observations, we assume the four observations on each individual to be independent observations on that individual. The dietary components included in the analyses are calcium, energy, iron, protein, vitamin A and vitamin C. These components were selected because of their nutritional importance and because of their different distributional behaviors.

Our model is

$$Z_{ij} = z_i + \epsilon_{ij}\,, \tag{3.8}$$

$$X_{ij} = g(Z_{ij};\, \theta)\,,$$

$$X_{ij} = x_i + u_{ij}\,,$$

$$\begin{bmatrix} x_i \\ u_{ij} \end{bmatrix} \sim \mathrm{NI}\left[\begin{bmatrix} \mu_x \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & 0 \\ 0 & \sigma_{uu} \end{bmatrix} \right],$$

where Z_{ij} is the observed intake for individual i on day j, $z_i = E\{Z_{\ell j} | \ell = i\}$ is the usual intake for individual i, and $g(Z_{ij}; \theta)$ is a transformation that maps observed intakes into normal random variables. In our research on this topic, considerable effort has been devoted to the specification and estimation of the function $g(Z_{ij}; \theta)$. For the purposes of our present discussion, we assume $g(Z_{ij}; \theta)$ is known.

To simplify our discussion, we treat the sample as a simple random sample of individuals. The variance components of the model can be estimated using the usual analysis of variance formulas. Thus,

$$\hat{\sigma}_{uu} = [n(r-1)]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{r} (X_{ij} - \bar{X}_{i.})^2 ,$$

$$\hat{\sigma}_{xx} = (n-1)^{-1} \sum_{i=1}^{n} (\bar{X}_{i.} - \bar{X}_{..})^2 - r^{-1}\hat{\sigma}_{uu} , \tag{3.9}$$

where $\bar{X}_{..} = n^{-1}\Sigma_{i=1}^{n}\bar{X}_{i.}$, $\bar{X}_{i.} = r^{-1}\Sigma_{j=1}^{r}X_{ij}$, and $r = 4$ is the number of replicate observations.

Table 4 contains the components of variance for the transformed data. The data have been transformed so that the transformed observations X_{ij} are approximately normally distributed with mean zero and variance one. The within variance exceeds the among variance for all dietary components. The ratio of within to among is smallest for energy with a value of 1.67 and is largest for vitamin A with a ratio of 2.92.

Figure 1 contains a plot of the original data against the normal scores for calcium. This plot is sometimes called a Q–Q plot. The line in the plot is the g–function. Figure 2 is the same plot for Z^{α}, where $\alpha = (3.5)^{-1}$. The root transformation produces variables that are nearly normal. The dashed lines in the plot are the join points for a grafted cubic fit to the empirical cumulative distribution function. Figure 3 is the normal score plot for protein, and Figure 4 is the corresponding plot for the square root of protein. For protein,

Table 4. Sample moments for data in the normal scale.

Dietary Component	Among–Individual Variance $\hat{\sigma}_x^2$	Within–Individual Variance $\hat{\sigma}_u^2$	$\frac{\hat{\sigma}_u^2}{\hat{\sigma}_x^2}$
Calcium	0.364	0.633	1.74
Energy	0.373	0.623	1.67
Iron	0.314	0.683	2.18
Protein	0.273	0.724	2.65
Vitamin A	0.254	0.742	2.92
Vitamin C	0.318	0.679	2.14

the root operation is not sufficient to produce normality and the grafted cubic completes the transformation to normality.

On the basis of Table 4, the true values (usual intakes) of calcium in the normal scale are distributed as $N(0, 0.633)$ random variables. The distribution of interest is the distribution of true values in the original scale. The true value in original scale for individual i is

$$z_i = E\{Z_{ij}|i\} = E\{g^{-1}(X_{ij})|i\} \tag{3.10}$$

$$= E\{g^{-1}(x + u)|x = x_i\} .$$

Because $g(\ )$ is a nonlinear function, the transformation that carries the distribution of x into the distribution of z is not g^{-1}. We can approximate the expectation for a particular x by numerical integration using $u \sim NI(0, \sigma_{uu})$. The calculated z_i for a set

of x_i are used to construct a smooth approximation to the relationship between x and z. Let

$$z_i = h(x_i) = E\{g^{-1}(x + u) | x = x_i\} . \tag{3.11}$$

Figure 5 contains a plot of the g and h functions for calcium. Because the g function is concave, the h function lies below the g function.

The estimated density for usual intakes in the original scale is

$$f_z(z) = \phi_x[h^{-1}(z)]\frac{\partial h^{-1}(z)}{\partial z} , \tag{3.12}$$

where $\phi_x(\ )$ is the distribution of usual intakes in normal space. Figure 6 contains the estimated density for the original observations, for the mean of four daily observations, and for usual intakes, all in the original scale for calcium. The original observations have a density with a relatively long right tail. The variance of the distribution of usual intakes is about 36% of the variance of one–day intakes and about 69% of the variance of a mean of four intakes per individual. The means of the three distributions are the same. The distribution of usual intakes is skewed, but less skewed than the distribution of one–day intakes. Figure 7 is the same plot for protein. The densities for protein are much more symmetric than those for calcium.

The estimation of the cumulative distribution function of the usual intakes in the presence of measurement error requires that at least two days be observed for some individuals in the sample. If one accepts the model, it is not necessary to have duplicate observations on every individual. The optimal design depends on the importance attached to the estimation of different portions of the distribution function and on the size of the error variances. Our analysis was prepared by Anthony An following suggestions of

Phil Kott. It is assumed that the analysis of variance estimators are used to estimate the response variance and the variance of the true values.

The entries in Table 5 are the variances for designs with different rates of replication under the assumption that independent identically distributed determinations are made on each individual. The first column of the table contains the number of observations per individual in the study. Thus, if one had 1000 observations and the entry is 1.25, there would be 800 individuals in the study and duplicate observations would be made on 200 individuals.

Because the sample mean is unbiased for the population mean under our model, the best design for the mean is to make no replicate determinations. If $\sigma_{uu} = 0.15$ and one is interested in the tails of the distribution, making duplicate observations on about 10 percent of the individuals is appropriate. If $\sigma_{uu} = 2.00$, a design with two determinations on each individual would be a good compromise design.

In reinterview studies, it is observed that the mean of the second interviews is not equal to the mean of the first interviews. See Bailar (1975). We call the differences in level observed for the different interviews, time–in–sample effects. One possible way to extend the model to include time–in–sample effects is to write

$$X_{ij} = x_i + \tau_j + u_{ij} , \tag{3.13}$$

where the τ_j are fixed time–in–sample effects. In the nutrition study, we used the first interview as the standard. For model (3.13), this is equivalent to setting τ_1 equal to zero. With fixed time–in–sample effects and $\tau_1 = 0$,

$$V\{\hat{\mu}\} = m^{-1}(\sigma_z^2 + \sigma_\epsilon^2) , \tag{3.14}$$

because $\hat{\mu}$ is the mean for the first interview. Table 5 was constructed using (3.7) with (3.3) as the variance of $\hat{\mu}$. Table 6 was constructed using (3.7) with (3.14) as the variance of $\hat{\mu}$. All variances in Table 6 are larger than those in Table 5. Also, the increase in variance of the estimated mean as the replication increases is more pronounced in Table 5 than in Table 5. For the case of $\sigma_{uu}/\sigma_{xx} = 2$, the design with two observations per individual is a good compromise design if the objective is to estimate the entire distribution function.

4. Estimation of Regression Equations

In this section, we consider estimation for the regression model

$$y_t = \beta_0 + \mathbf{x}_t\beta_1 + q_t , \tag{4.1}$$

where

$$\mathbf{Z}_t = (Y_t, \mathbf{X}_t) = (y_t, \mathbf{x}_t) + (w_t, \mathbf{u}_t) = \mathbf{z}_t + \epsilon_t$$

$$\mathbf{x}_t \sim \text{Ind}(\mu_x, \Sigma_{xx}) ,$$

and

$$\epsilon_t \sim \text{Ind}(0, \Sigma_{\epsilon\epsilon}) .$$

The q_t is called the error in the equation. It is assumed that

$$E\{\epsilon_t | \mathbf{x}_t\} = 0$$

$$E\{q_j\epsilon_t\} = 0$$

Table 5. Variances of estimated normal quantiles with identically distributed replicates standardized to one observation and $\sigma_{xx} = 1.0$.

Obs. per individual	Q(50)	Q(75)	Q(90)	Q(95)	Q(97.5)	Q(99)
			$\sigma_{uu} = 0.15$			
1.05	1.20	1.61	2.68	3.64	4.66	6.07
1.10	1.26	1.63	2.59	3.46	4.38	5.66
1.25	1.41	1.79	2.76	3.64	4.57	5.85
1.50	1.67	2.09	3.19	4.17	5.22	6.67
2.00	2.15	2.68	4.06	5.30	6.62	8.45
3.00	3.15	3.91	5.88	7.64	9.53	12.14
			$\sigma_{uu} = 1.00$			
1.05	2.07	7.28	20.87	33.04	46.04	64.01
1.10	2.15	5.12	12.90	19.86	27.30	37.58
1.25	2.34	3.99	8.30	12.16	16.28	21.98
1.50	2.62	3.88	7.15	10.08	13.20	17.53
2.00	3.00	4.14	7.11	9.78	12.62	16.55
3.00	4.00	5.25	8.53	11.46	14.59	18.93
			$\sigma_{uu} = 2.00$			
1.05	3.10	22.30	72.41	117.27	165.21	231.48
1.10	3.19	13.44	40.20	64.16	89.76	125.15
1.25	3.44	8.27	20.90	32.21	44.28	60.98
1.50	3.75	6.82	14.84	22.03	29.71	40.32
2.00	4.00	6.28	12.23	17.55	23.24	31.10
3.00	5.00	7.05	12.41	17.21	22.34	29.42
			$\sigma_{uu} = 3.50$			
1.50	5.44	12.95	32.56	50.13	68.89	94.83
2.00	5.50	10.34	22.98	34.29	46.37	63.08
3.00	6.50	10.18	19.78	28.37	37.55	50.25

Table 6. Variances of estimated normal quantiles with time–in–sample effects standardized to one observation and $\sigma_{xx} = 1.0$.

Obs. per individual	Q(50)	Q(75)	Q(90)	Q(95)	Q(97.5)	Q(99)
			$\sigma_{uu} = 0.15$			
1.05	1.21	1.62	2.68	3.64	4.66	6.07
1.10	1.27	1.64	2.60	3.47	4.39	5.67
1.25	1.44	1.81	2.79	3.66	4.59	5.88
1.50	1.72	2.15	3.24	4.22	5.27	6.72
2.00	2.30	2.83	4.21	5.45	6.77	8.60
3.00	3.45	4.21	6.18	7.94	9.83	12.44
			$\sigma_{uu} = 1.00$			
1.05	2.10	7.31	20.90	33.07	46.07	64.04
1.10	2.20	5.18	12.95	19.91	27.35	37.64
1.25	2.50	4.15	8.46	12.32	16.44	22.14
1.50	3.00	4.25	7.52	10.45	13.58	17.90
2.00	4.00	5.14	8.11	10.78	13.62	17.55
3.00	6.00	7.25	10.53	13.46	16.59	20.93
			$\sigma_{uu} = 2.00$			
1.05	3.15	22.35	72.46	117.33	165.26	231.54
1.10	3.30	13.55	40.31	64.27	89.87	125.26
1.25	3.75	8.59	21.21	32.52	44.60	61.29
1.50	4.50	7.57	15.60	22.78	30.46	41.07
2.00	6.00	8.28	14.23	19.55	25.24	33.10
3.00	9.00	11.05	16.41	21.21	26.34	33.42
			$\sigma_{uu} = 3.50$			
1.50	6.75	14.26	33.88	51.44	70.20	96.14
2.00	9.00	13.84	26.48	37.79	49.87	66.58
3.00	13.50	17.18	26.78	35.37	44.55	57.25

for all t and all j. If replicate observations are made on the ϵ_t, we denote the replicate observations by ϵ_{tj}. We assume $|\Sigma_{xx}| > 0$ and $\sigma_{qq} > 0$.

The sum of $q_t + w_t$ is sometimes denoted by e_t. Given a sample of n observations, the ordinary least squares estimator of β_1 is

$$\hat{\gamma}_1 = \mathbf{m}_{XX}^{-1}\mathbf{m}_{XY}, \tag{4.2}$$

where

$$\mathbf{m}_{XX} = (n-1)^{-1} \sum_{t=1}^{n} (X_t - \bar{X})'(X_t - \bar{X}),$$

$$\mathbf{m}_{XY} = (n-1)^{-1} \sum_{t=1}^{n} (X_t - \bar{X})'(Y_t - \bar{Y}),$$

$$\bar{Z} = (\bar{Y}, \bar{X}) = n^{-1} \sum_{t=1}^{n} (Y_t, X_t).$$

From equation (2.9), the expected values are

$$E\{\mathbf{m}_{XX}, \mathbf{m}_{XY}\} = (\Sigma_{xx} + \Sigma_{uu}, \Sigma_{xy} + \Sigma_{uw}). \tag{4.3}$$

If (x_t, ϵ_t, q_t) are independent normal vectors, then

$$E\{\hat{\gamma}_1\} = \Sigma_{XX}^{-1}\Sigma_{XY}. \tag{4.4}$$

If $\Sigma_{uu} \neq 0$, the ordinary least squares estimator is biased for β_1 unless

$$\Sigma_{uu}\beta_1 = \Sigma_{uw}. \tag{4.5}$$

In many situations, it is reasonable to assume that $\Sigma_{uw} = 0$. In such cases, $\hat{\gamma}_1$ is a biased estimator for β_1 in the presence of measurement error. The magnitude of the bias depends on the magnitude of the measurement error variance relative to the variance of the true values. For our model with a single explanatory variable and $\sigma_{uw} = 0$,

$$\gamma_1 = E\{\hat{\gamma}_1\} = \kappa_{xx}\beta_1 , \tag{4.6}$$

where $\kappa_{xx} = \sigma_{XX}^{-1}\sigma_{xx}$.

The ratio κ_{xx} is sometimes called the reliability ratio. Fuller (1987, p. 8) gives values of κ_{xx} for socioeconomic variables often collected in surveys. Some of the values are 0.98, 0.92, 0.88, 0.85, and 0.77 for sex, the 45–49 age category, education, income, and unemployment status, respectively. Thus, the ordinary simple least squares regression coefficient for income is biased by about 15% of its true value. For the zero–one variables such as sex, the reported ratio is the ratio for the latent class model, the model that defines the true values so that the mean of the response error is zero for every individual.

For normal variables with $n > 3$,

$$V\{\hat{\gamma}_1\} = (n-3)^{-1}\sigma_{XX}^{-1}(\sigma_{YY} - \gamma_1\sigma_{XY}) . \tag{4.7}$$

The mean square error of $\hat{\gamma}_1$ as an estimator of β_1 is

$$E\{(\hat{\gamma}_1 - \beta_1)^2\} = (\kappa_{xx} - 1)^2\beta_1^2 + V\{\hat{\gamma}_1\} . \tag{4.8}$$

A sample of observations Z_t, $t = 1, 2, ..., n$ is not sufficient for the estimation of the parameters of model (4.1). Some type of additional information is required. Several sources can be considered. One possible approach is to use instrumental variables, where instrumental variables are variables correlated with x_t but not correlated with ϵ_t. A

related approach is to develop an augmented model in which relationships among a number of variables is specified. See, for example, Fuller (1987), for discussions of such techniques.

For our current discussion, we restrict consideration to the procedure of using replicate determinations to estimate $\Sigma_{\epsilon\epsilon}$. Assume that it is possible to make independent identically distributed observations on some elements of interest. Let Z_{tj}, $j = 1, 2,$ be two determinations on element t. If two independent identical determinations are made on each of d elements,

$$\hat{\Sigma}_{\epsilon\epsilon} = 0.5\, d^{-1} \sum_{t=1}^{d} (Z_{t1} - Z_{t2})'(Z_{t1} - Z_{t2}) \tag{4.9}$$

is an unbiased estimator of $\Sigma_{\epsilon\epsilon}$. Let $\bar{Z}_{t.} = 0.5(Z_{t1} + Z_{t2})$ and $\bar{Z}_{..} = d^{-1}\Sigma_{t=1}^{d}\bar{Z}_{t.}$. Then

$$\hat{\Sigma}_{zz} = (d-1)^{-1} \sum_{t=1}^{d} (\bar{Z}_{t.} - \bar{Z}_{..})'(\bar{Z}_{t.} - \bar{Z}_{..}) - 0.5\hat{\Sigma}_{\epsilon\epsilon} \tag{4.10}$$

is an unbiased estimator of Σ_{zz}. It follows that consistent estimators are

$$\hat{\beta}_1 = \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}, \tag{4.11}$$

and

$$\hat{\beta}_0 = \bar{Y}_{..} - \bar{X}_{..}\hat{\beta}_1, \tag{4.12}$$

where

$$\hat{\Sigma}_{zz} = \begin{bmatrix} \hat{\sigma}_{yy} & \hat{\Sigma}_{yx} \\ \hat{\Sigma}_{xy} & \hat{\Sigma}_{xx} \end{bmatrix}.$$

In practice, it is necessary to check that $\hat{\Sigma}_{zz}$ is positive semidefinite before constructing the estimator of β_1. See Fuller (1987, p. 105) for the construction of estimators when $\hat{\Sigma}_{zz}$ is not positive definite. It is possible to show that $d^{1/2}(\hat{\beta}_1 - \beta_1)$ has a limiting normal distribution with mean zero and covariance matrix

$$\Gamma = \Sigma_{xx}^{-1}\sigma_{vv} + \Sigma_{xx}^{-1}[\Sigma_{uu}(\sigma_{vv} + \sigma_{rr}) + 2\Sigma_{uv}\Sigma_{vu}]\Sigma_{xx}^{-1}, \tag{4.13}$$

where

$$\sigma_{vv} = (1, -\beta_1')\Sigma_{ZZ}(1, -\beta_1')' ,$$

$$\sigma_{rr} = (1, -\beta_1')\Sigma_{\epsilon\epsilon}(1, -\beta_1')' ,$$

$$\Sigma_{uv} = \Sigma_{u\epsilon}(1, -\beta_1')' .$$

It is possible to extend the estimation procedure to the situation in which a different number of replications are made on different elements of the sample. The analysis of variance estimators are not fully efficient with unequal numbers of replications. Fuller (1991) and Sanger (1992) give efficient estimation procedures.

To construct the measurement error estimator, one must have available an estimator of $\Sigma_{\epsilon\epsilon}$. If a study has not been previously conducted, some of the resources from the current study must be used to estimate $\Sigma_{\epsilon\epsilon}$. In such a situation, the investigator must choose between a design with no replication and the ordinary least squares estimator and a design that allocates resources to the estimation of $\Sigma_{\epsilon\epsilon}$.

If a sample is very small, the mean square error of the ordinary least squares estimator will be smaller than the variance of the estimator adjusted for measurement error. Because the variance of the measurement error estimator is a function of Σ_{zz} and

Table 7. Sample size for which variance of MEM with $\delta = 0.25$ equals the MSE of OLS.

κ_{xx}	R^2_{xy}		
	0.25	0.50	0.75
0.98	3274	1320	662
0.92	330	168	109
0.88	189	103	70
0.85	143	80	56
0.77	88	53	38
0.58	55	35	26

$\Sigma_{\epsilon\epsilon}$, the optimal design is a function of unknown parameters. Table 7 compares two design–estimation procedures for the simple regression problem. One procedure is to observe n elements and use the ordinary least squares estimator. The second procedure observes 0.75n different individuals, makes duplicate observations on 0.25n of the 0.75n individuals, uses the duplicate observations to estimate σ_{uu} and uses the estimator given by Fuller (1991) to estimate β_1. It is assumed that σ_{uw} is known to be zero. The numbers in the table are the sample sizes at which the mean square error of the ordinary least squares estimator is equal to the variance of the measurement error estimator. Thus, if the ratio of σ_{xx} to σ_{XX} is 0.85 and if the squared correlation between y and true x is 0.25, the procedure that uses 25% of the observations as replicate observations to estimate the error variance is superior to the ordinary least squares procedure if one is able to make 143 or more determinations. In the case of the measurement error procedure, 107 different individuals would be included in the study and 36 individuals would be observed twice. Thus, for measurement error variances of the magnitude recorded for socioeconomic surveys and for the mean square error criterion, resources should be allocated to error

variance estimation when the survey is to be used for regression estimation and sample size exceeds 150.

We have considered only simple regression. The effects of measurement error generally increase as the size of the regression problem increases. See Fuller (1991) and references cited there.

5. Comments

We have demonstrated that sample quantiles and regression coefficients are biased in the presence of measurement error. We were able to construct consistent estimators for these parameters using estimators of the error variances of the measurement error.

In our discussion, we have restricted our attention to simple models with strong assumptions. A number of extensions appear in the literature. See Stefanski and Carroll (1990, 1991) and references cited there for discussions of the estimation of the density function for a variable contaminated with measurement error. Fuller (1987, 1991) discusses a number of extensions for regression estimation, including extensions to models with unequal error variances. The assumption of normal distributions is not required for most regression estimation and was used in our presentation only for convenience. However, as with many other procedures, the estimators can be heavily influenced by extreme observations. Also, in some survey situations, the distribution of the measurement error may have heavy tails. There is limited work on robust procedures for measurement error problems.

A primary concern of Hansen and his co–workers at the U.S. Census was with the correlation between responses on different individuals. See Hansen et al. (1951, 1961, 1964). Bailar (1983) contains a discussion of methods of estimating this correlation. The estimation techniques we have described can be extended to the case of correlated measurement errors, where the correlation is between observations on different individuals.

There are numerous practical considerations associated with the implementation of replicate determinations on questionnaire items for human subjects. Some of these are discussed in our earlier references. Also see Bailar (1968) and Forsman and Schreiner (1991).

We assumed it is possible to make independent determinations on the same individual. In practice, there are two competing considerations. One desires the determinations to be close together in time to ensure that the respondent is responding for the same item and also that the respondent is available for contact. On the other hand, responses made close together in time are more likely to be correlated than those more separated in time.

Table 8 has been constructed to evaluate the effect of correlated individual determinations on the estimator defined by (4.11) and (4.10). We assume that the estimator is constructed using replicated observations acting as if the replicates are independent. We assume that a total of n determinations are made and that $\delta n = d$ of these are replicate observations, where $\delta \leq 0.5$. Thus, $n - d = n(1 - \delta)$ individuals are observed, of whom $n(1 - 2\delta)$ are observed only once. We assume that it is known that $\sigma_{uw} = 0$ and, hence, only the error variance in X need be estimated.

The effect of the correlation on the estimators increases as the size of the measurement variance increases and as the size of the sample increases. For $\kappa_{xx} = 0.85$, the reliability ratio for income, there is about a 10% loss in efficiency with a correlation of 0.10 relative to independent observations at a sample size of 1000. If the correlation is 0.25, the squared bias at $n = 5000$ is about four times the variance of the measurement error procedure. On the other hand, the procedure of making no determinations and using the ordinary least squares estimator has a mean square error about 15 times that of the biased measurement error procedure.

We focused on estimators in two situations, quantile estimation and linear regression. See Carroll (1992) and Fuller (1987) for discussions of estimation for nonlinear

models. Also, we have concentrated on one form of information about the measurement error distribution, that obtained from replicate observations. Other procedures such as those based on instrumental variables are available. See, for example, Fuller (1987).

Hopefully, our discussion will stimulate the use of existing methodology and research on procedures tailored for survey samples. Our models were relatively simple. Extensions of the methodology to combine several kinds of information about the measurement error, extensions to models where the error distribution is a function of observables, variance estimation for estimators constructed from complex surveys and development of procedures robust against extreme observations are but a few of the areas deserving research attention.

Table 8. Mean square error multiplied by n for alternative estimation schemes and correlation between replicated determinations $\sigma_{xx} = 1$, $\beta_1 = 1$, $R^2_{xy} = 0.50$.

n	$\delta = 0.25$			OLS
	$\rho = 0$	$\rho = 0.10$	$\rho = 0.25$	
	$\sigma_{uu} = 0.087 (\kappa_{xx} = 0.92)$			
100	1.58	1.56	1.55	1.63
200	1.58	1.56	1.59	2.27
500	1.58	1.59	1.73	4.20
1000	1.58	1.62	1.95	7.40
5000	1.58	1.92	3.77	33.02
	$\sigma_{uu} = 0.1765 (\kappa_{xx} = 0.85)$			
100	1.96	1.88	1.89	3.23
200	1.96	1.91	2.07	5.48
500	1.96	2.00	2.61	12.23
1000	1.96	2.15	3.50	23.48
5000	1.96	3.36	10.64	113.51
	$\sigma_{uu} = 0.2987 (\kappa_{xx} = 0.77)$			
100	2.67	2.46	2.53	6.24
200	2.67	2.55	3.01	11.53
500	2.67	2.80	4.46	27.40
1000	2.67	3.22	6.87	53.85
5000	2.67	6.59	26.18	265.45

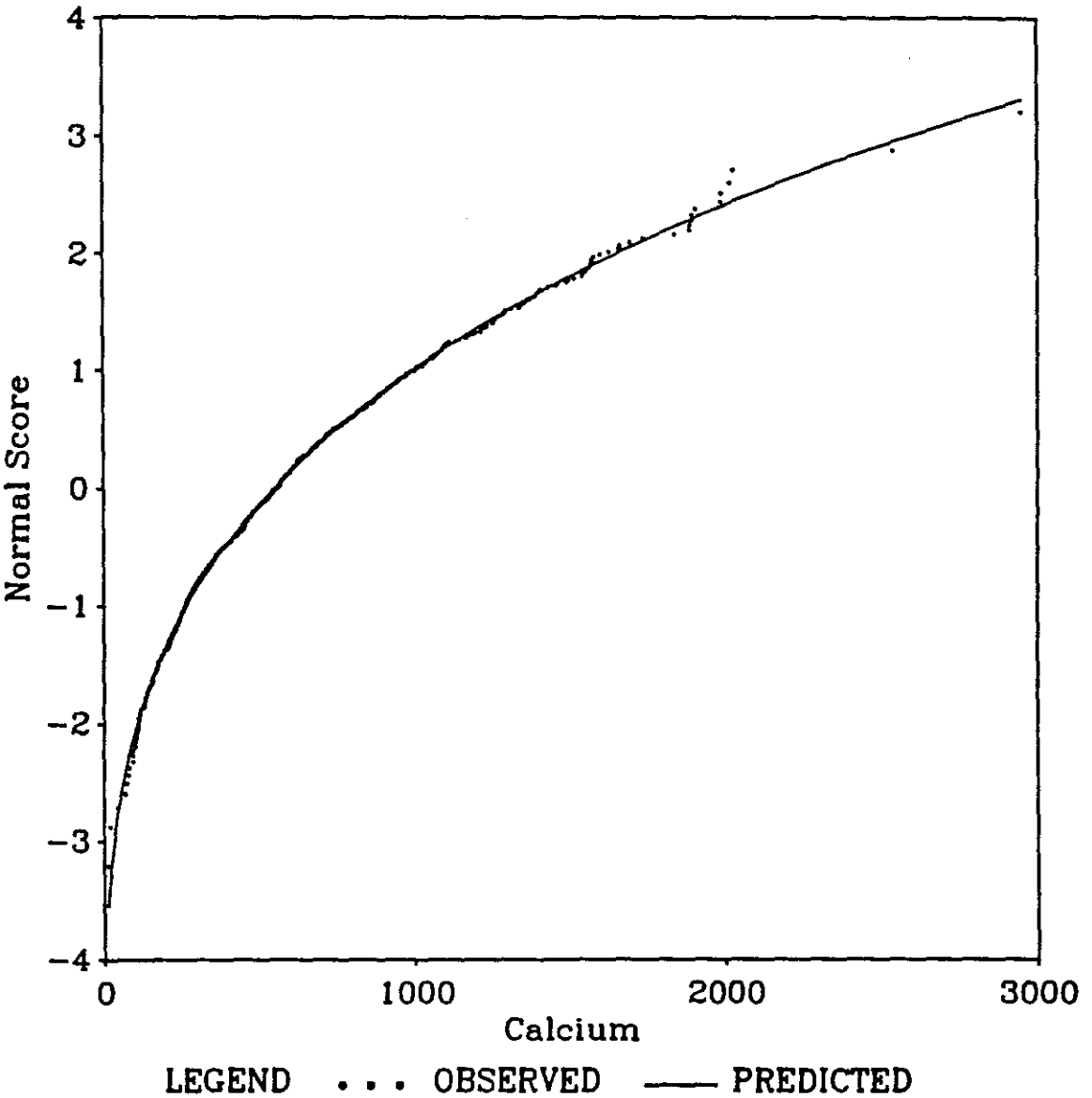

Figure 1. Normal score plot for calcium

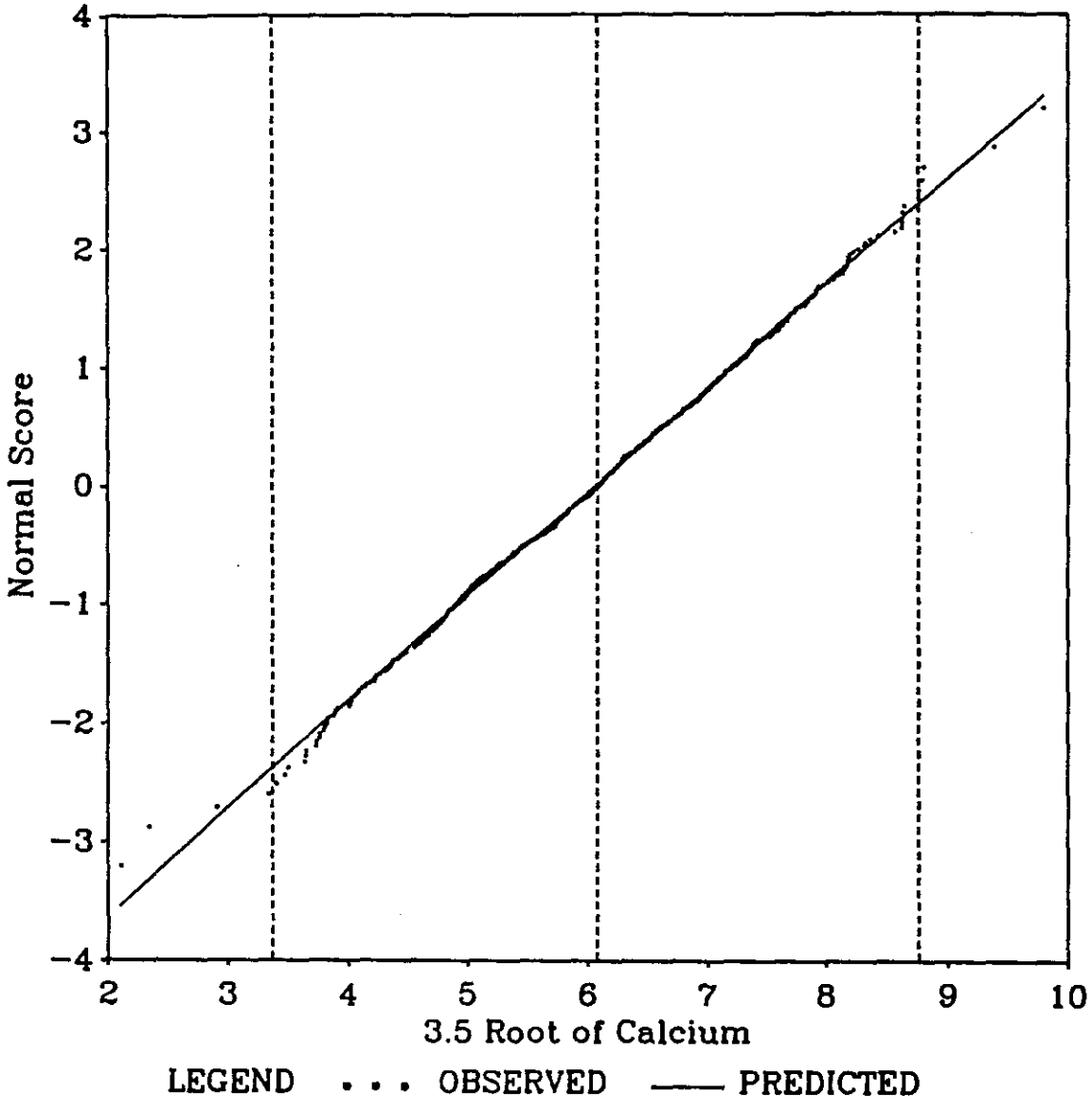

Figure 2. Normal score plot for the $(3.5)^{-1}$ power of calcium

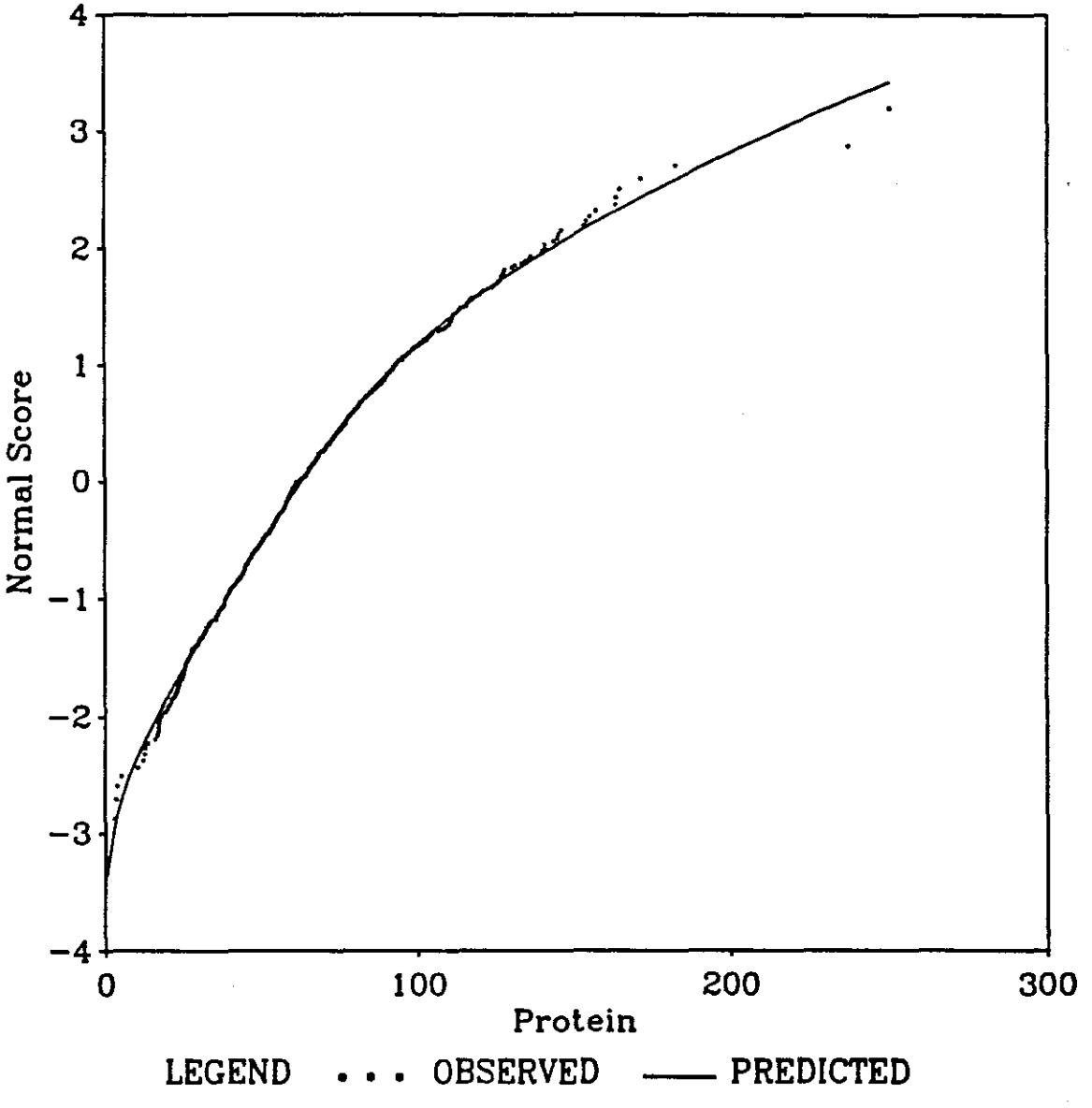

Figure 3. Normal score plot for protein

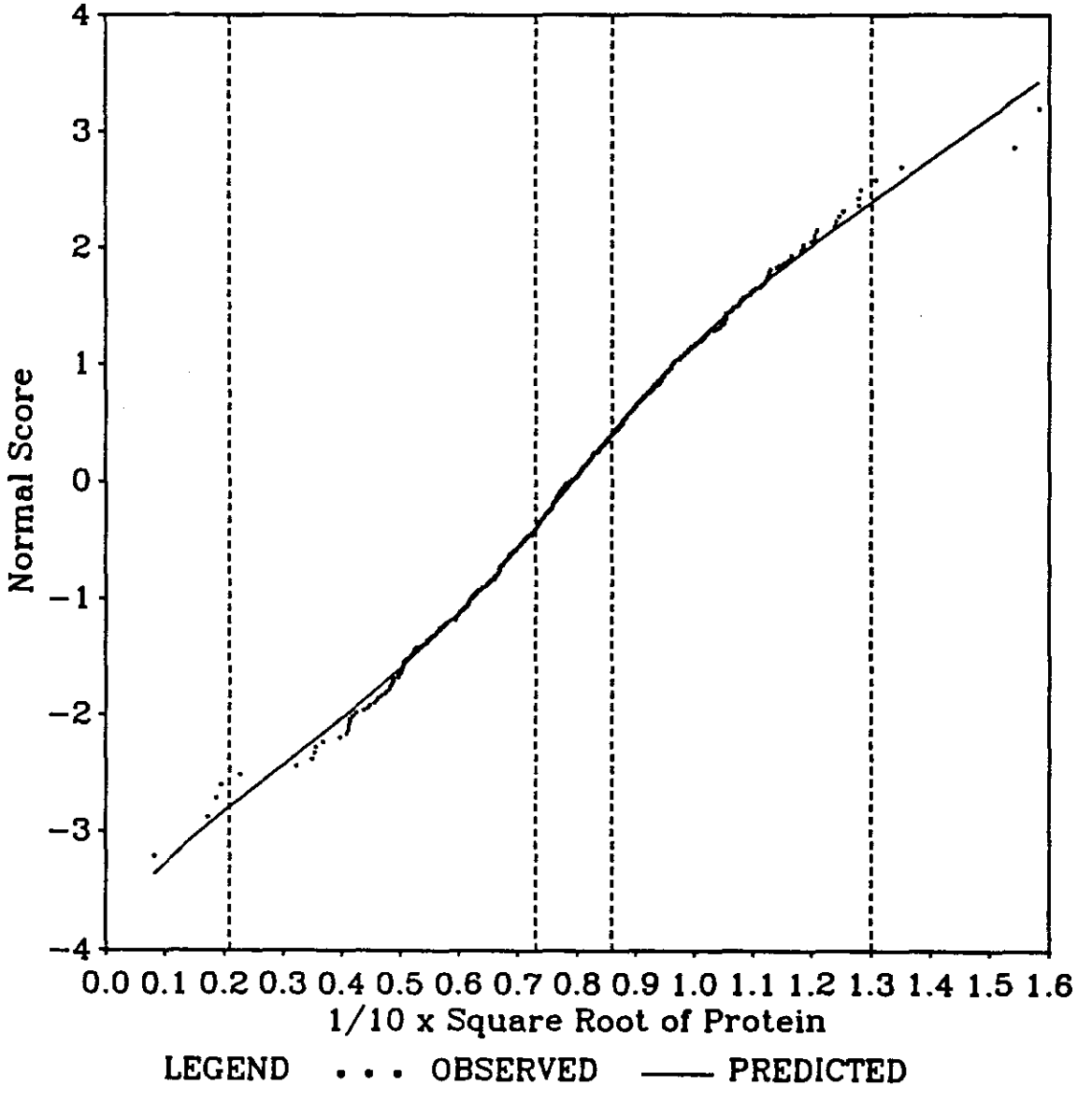

Figure 4. Normal score plot for square root of protein

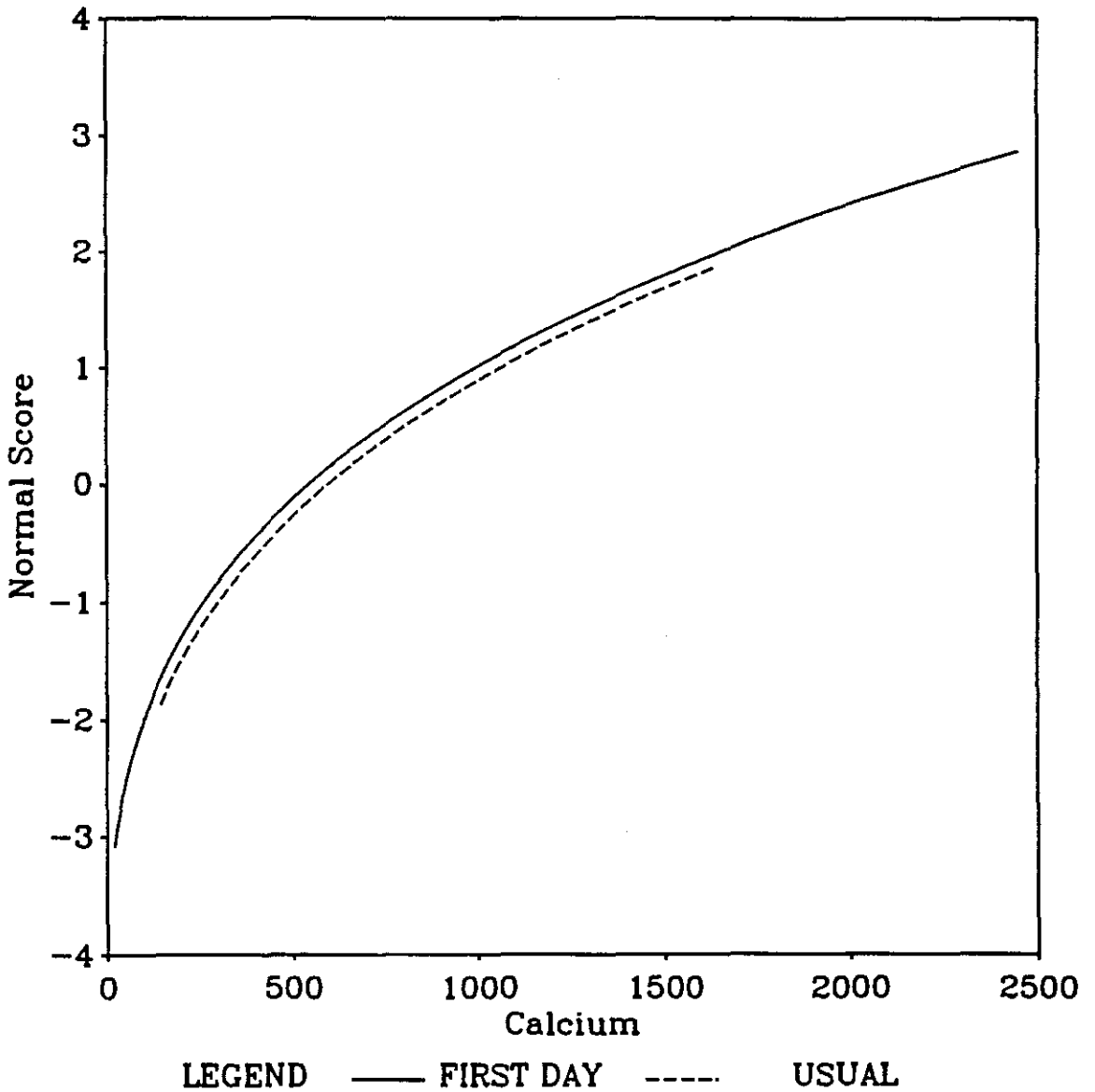

Figure 5. The g and h functions for calcium

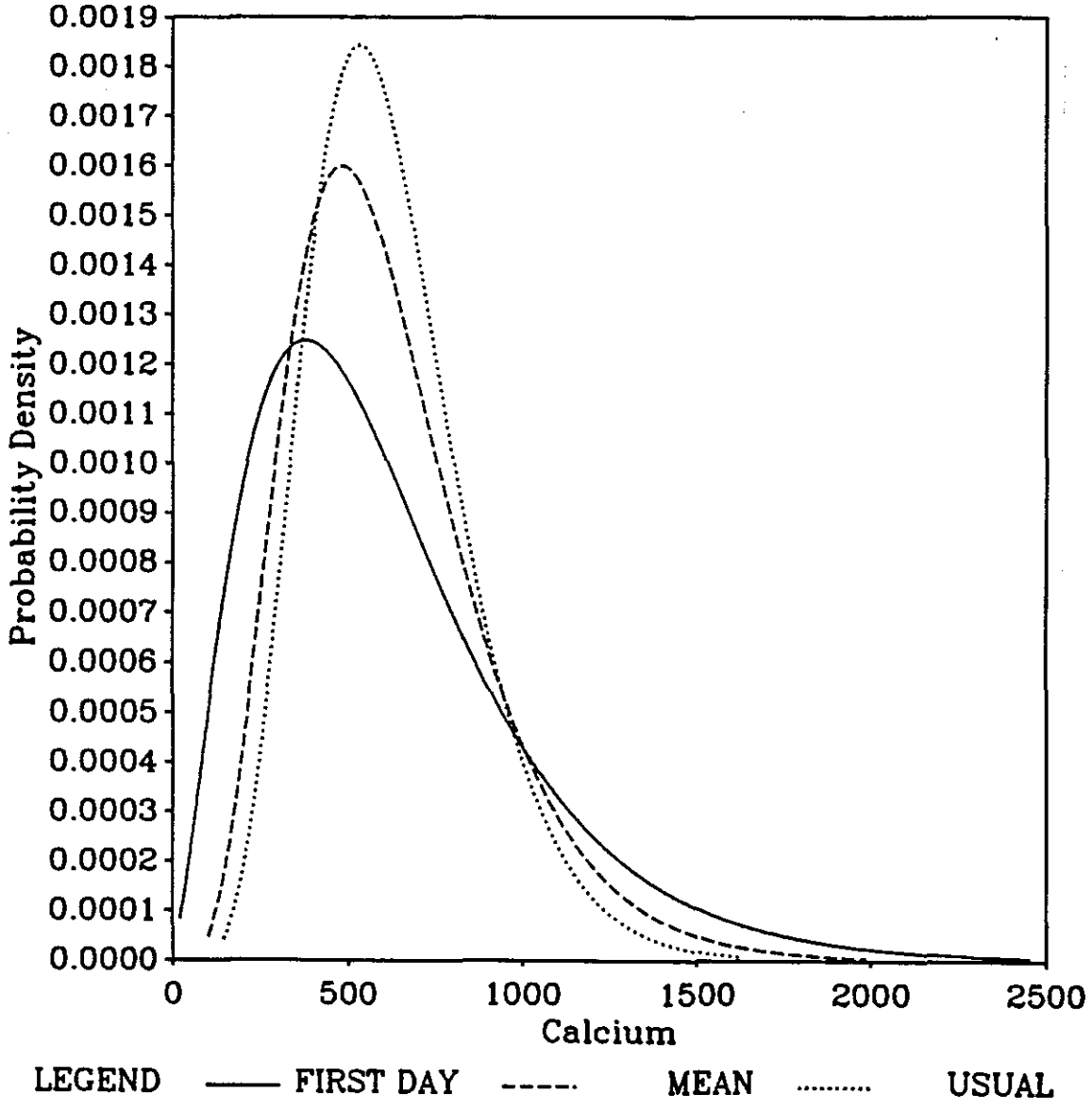

Figure 6. Estimated densities for calcium

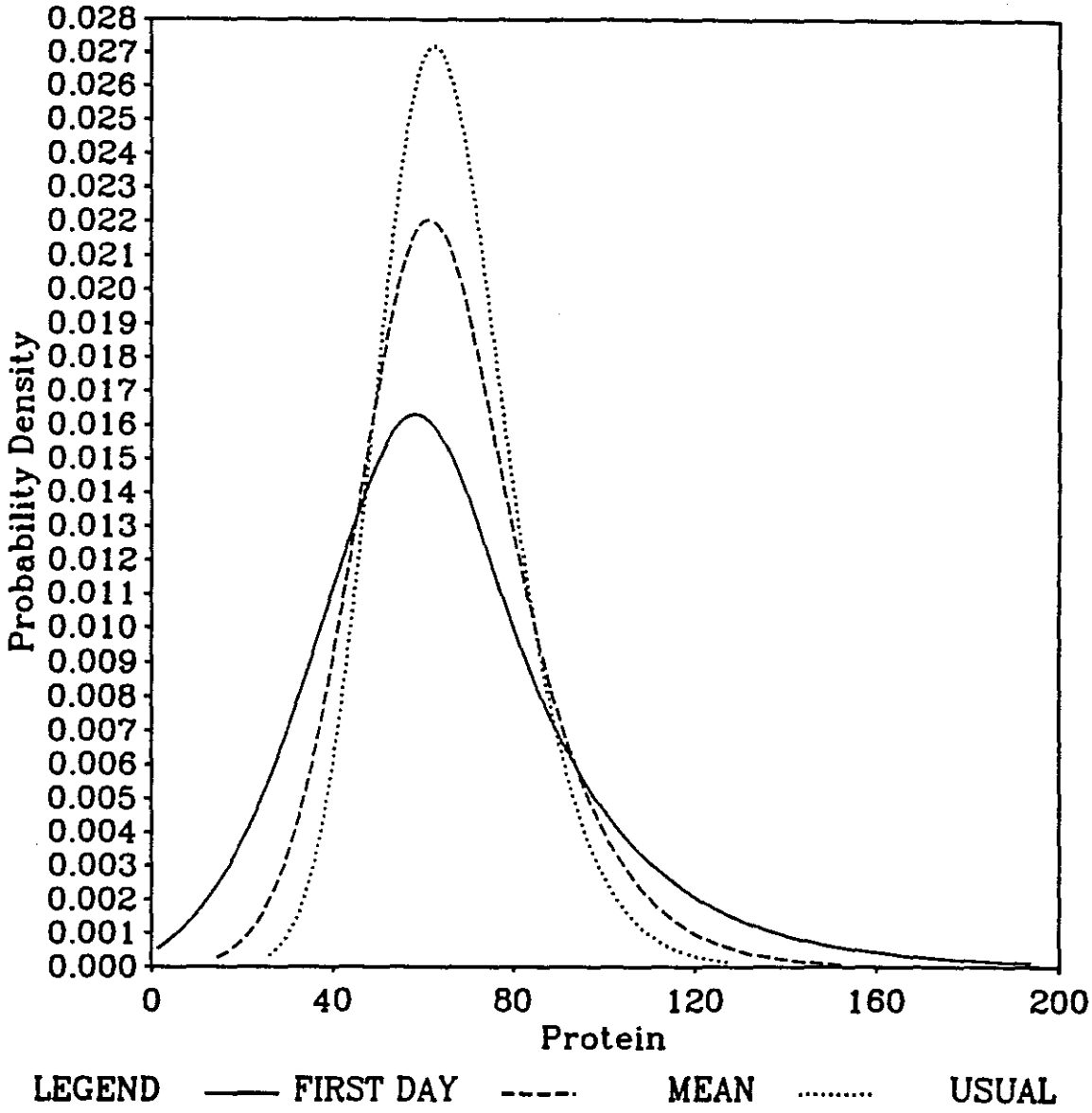

Figure 7. Estimated densities for protein

APPENDIX A.

In this appendix, we give the results used in constructing Table 8. The error variance is estimated using replicate observations on X. If there is a correlation of ρ_u between the two observations on X_t, then

$$E\{\hat{\sigma}_{uu}\} = \sigma_{uu}(1 - \rho_u)$$

and

$$V\{0.5(X_{t1} + X_{t2}) - x_t\} = 0.5\sigma_{uu}(1 + \rho_u) ,$$

where

$$\hat{\sigma}_{uu} = d^{-1} \sum_{t=1}^{d} (X_{t1} - X_{t2})^2 .$$

We assume that the covariance matrix of the true values is estimated with

$$\hat{\Sigma}_{zz} = (n - d)^{-1} \sum_{t=1}^{n-d} (\bar{Z}_{t.} - \bar{Z}_{..})'(\bar{Z}_{t.} - \bar{Z}_{..}) - (1 - \delta)^{-1}(1 - 1.5\delta)\text{diag}(0, \hat{\sigma}_{uu}) ,$$

where $\bar{Z}_{t.} = 0.5(Z_{t1} + Z_{t2})$ for $t = 1, 2, ..., d$, and $\bar{Z}_{t.} = Z_{t1}$ for $t = d + 1, d + 2, ..., n - d$, and

$$\bar{Z}_{..} = (n - d)^{-1} \sum_{t=1}^{n-d} \bar{Z}_{t.} .$$

Therefore, the approximate expected value of the estimator

$$\hat{\beta}_1 = \hat{\sigma}_{xx}^{-1}\hat{\sigma}_{xy}$$

is

$$E\{\hat{\beta}_1\} = (\sigma_{xx} + \rho_u \sigma_{uu})^{-1} \sigma_{xx} \beta_1 .$$

Assume there is no measurement error in y . Then the covariance matrix of $(\hat{\sigma}_{xy}, \hat{\sigma}_{xx})$ is

$$\mathbf{V}_{\sigma\sigma} = n^{-1}(1-\delta)^{-2}[\delta \mathbf{V}_{11} + (1-2\delta)\mathbf{V}_{22} + (1-1.5\delta)^2 \delta^{-1} \mathbf{V}_{33}] ,$$

where

$$\mathbf{V}_{11} = \begin{bmatrix} \sigma_{YY}\sigma_{XX11} + \sigma_{XY}^2 & 2\sigma_{XX11}^2 \sigma_{XY} \\ 2\sigma_{XX11}\sigma_{XY} & 2\sigma_{XX11}^2 \end{bmatrix} ,$$

$$\mathbf{V}_{22} = \begin{bmatrix} \sigma_{YY}\sigma_{XX} + \sigma_{XY}^2 & 2\sigma_{XX}\sigma_{XY} \\ 2\sigma_{XX}\sigma_{XY} & 2\sigma_{XX}^2 \end{bmatrix} ,$$

$$\mathbf{V}_{33} = \begin{bmatrix} 0 & 0 \\ 0 & 2[\sigma_{uu}(1-\rho)]^2 \end{bmatrix} ,$$

and $\sigma_{XX11} = \sigma_{xx} + 0.50\sigma_{uu}(1+\rho)$. Let

$$C = E\{\hat{\beta}_1\} ,$$

and

$$nV\{\hat{\beta}_1\} = (\sigma_{xx} + \rho\sigma_{uu})^{-2}(1, -C)\mathbf{V}_{\sigma\sigma}(1, -C)' .$$

Then the mean square error of $\hat{\beta}_1$ as an estimator of β_1 is

$$E\{(\hat{\beta}_1 - \beta_1)^2\} \doteq (\beta_1 - C)^2 + V\{\hat{\beta}_1\} .$$

REFERENCES

Bailar, B. A. (1968). Recent research in reinterview procedures. J. Amer. Statist. Assoc. **63** 41–63.

Bailar, B. A. (1975). The effects of rotation group bias on estimates from panel surveys. J. Amer. Statist. Assoc. **70** 23–29.

Bailar, B. A. (1983). Interpenetrating subsamples. In Johnson, N. L. and Kotz, S. (eds.) Encyclopedia of Statistical Sciences 4 197–201.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (eds.) (1991). Measurement Errors in Surveys. Wiley, New York.

Carroll, R. J. (1992). Approaches to estimation with errors in predictors. In Fahrmeir, L., Francis, B., Gilchrist, R., and Tutz, G. (eds.) Advances in GLIM and Statistical Modelling. Springer–Verlag, New York.

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

Forsman, G. and Schreiner, I. (1991). The design and analysis of reinterview: an overview. In Biemer, P. P. et al. (eds.) Measurement Errors in Surveys. Wiley, New York.

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

Fuller, W. A. (1991). Regression estimation in the presence of measurement error. In Biemer, P. P. et al. (eds.) Measurement Errors in Surveys. Wiley, New York.

Hansen, M. H., Hurwitz, W. N., and Bershad, M. A. (1961). Measurement errors in censuses and surveys. Bulletin of the International Statistical Institute, **38** 359–374.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). Sample Survey Methods and Theory. Wiley, New York. Vols I and II.

Hansen, M. H., Hurwitz, W. N., Marks, E. S., and Mauldin, W. P. (1951). Response errors in surveys. J. Amer. Statist. Assoc. **46** 147–190.

Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance. Contributions to Statistics Presented to Professor P. C. Mahalanobis on the Occasion of His 70th Birthday. Pergamon Press, Calcutta, India.

Lessler, J. T. and Kalsbeck, W. D. (1992). Nonsampling Error in Surveys. Wiley, New York.

Nusser, S. M., Carriquiry, A. L., Dodd, K. and Fuller, W. A. (1992). A semiparametric transformation approach to estimating usual intake distributions. J. American Statistical Association. To appear.

Sanger, T. M. (1992). Estimated generalized least squares estimation for the heterogeneous measurement error model. Ph.D. dissertation. Iowa State University, Ames, Iowa.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. Statistics, **21** 169–184.

Stefanski, L. A. and Carroll, R. J. (1991). Deconvolution–based score tests in measurement error models. Ann. Statist. **19** 249–259.

U.S. Department of Commerce (1975), 1970 Census of Population and Housing. Accuracy of Data for Selected Population Characteristics as measured by the 1970 CPS–Census Match. PHE(E)–11 U.S. Government Printing Office, Washington, D.C.