

# DESIGN ASPECTS OF CALIBRATION STUDIES IN NUTRITION, WITH ANALYSIS OF MISSING DATA IN LINEAR MEASUREMENT ERROR MODELS

Raymond J. Carroll, Laurence Freedman and David Pee \*

January 13, 1997

## Abstract

Motivated by an example in nutritional epidemiology, we investigate some design and analysis aspects of linear measurement error models with missing surrogate data. The specific problem investigated consists of an initial large sample in which the response (a food frequency questionnaire, FFQ) is observed, and then a smaller calibration study in which replicates of the error prone predictor are observed (food records or recalls, FR). The difference between our analysis and most of the measurement error model literature is that in our study, the selection into the calibration study can depend upon the value of the response. Rationale for this type of design is given. Two major problems are investigated. In the design of a calibration study, one has the option of larger sample sizes and fewer replicates, or smaller sample sizes and more replicates. Somewhat surprisingly, neither strategy is uniformly preferable in cases of practical interest. The answers depend on the instrument used (recalls or records) and the parameters of interest. The second problem investigated is one of analysis. In the usual linear model with no missing data, method of moments estimates and normal-theory maximum likelihood estimates are approximately equivalent, with the former method in most use because it can be calculated easily and explicitly. Both estimates are valid without any distributional assumptions. In contrast, in the missing data problem under consideration, only the moments estimate is distribution-free, but the maximum likelihood estimate has at least 50% greater precision in practical situations when normality obtains. Implications for the design of nutritional calibration studies are discussed.

*Key words and phrases:* Sampling Designs; Errors-in-Variables; Estimating Equations; Linear regression; Maximum Likelihood; Measurement Error; Method of Moments; Missing Data; Model Robustness; Nutrition; Semiparametrics; Stratified Sampling; Weighting.

**Short title.** Calibration Studies in Nutrition.

---

\*Raymond J. Carroll is Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Laurence Freedman is with the Biometry Branch, DCPC, National Cancer Institute, Executive Plaza North, Room 344, MSC 7354, Bethesda, MD 20892-7354. David Pee is with Information Management Services, Inc., 6120 Executive Boulevard, Rockville, MD 20852. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030). Carroll's research was partially completed while visiting the Institut für Statistik und Ökonometrie, Sonderforschungsbereich 373, Humboldt Universität zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award.

# 1 INTRODUCTION

## 1.1 Overview

The assessment and quantification of an individual's usual diet is a difficult exercise, but one that is fundamental to discovering relationships between diet and disease and to monitoring dietary behavior among individuals and populations. Various dietary assessment instruments have been devised, of which three main types are most commonly used in contemporary nutritional research. The one that is most convenient and inexpensive to use is the Food Frequency Questionnaire (FFQ), which is the instrument of choice in large nutritional epidemiology studies. However, while dietary intake levels reported from FFQ's are correlated with true usual intake, they are thought to involve a systematic bias (i.e. under-or-over-reporting at the level of the individual). The other two instruments that are commonly used are the 24-hour food recall and the multiple-day food record (FR). Each of these is more work- intensive and more costly, but is thought to involve less bias than a FFQ. However, the large daily variation in a western diet makes a single FR an imprecise measure of true usual intake.

## 1.2 Calibration Studies and Their Aims

Despite the problem that FR's are not completely unbiased, and may involve some under-reporting, their generally accepted superiority over FFQ's makes them the current practical gold- standard for dietary assessment. Thus, for proper interpretation of epidemiologic studies that use FFQ's as the basic dietary instrument, one needs to know the relationship between reported intakes from the FFQ and true usual intake, defined by the average intakes reported over a very long series of FR's. Such a relationship is ascertained through a substudy, commonly called a calibration (or validation) study.

The design of calibration studies has only recently attracted the interest of biostatisticians. This interest arises from the growing awareness of the problem of error in the measurement of exposures and its effect on estimation and power in epidemiologic studies. Calibration studies can provide valuable information on the nature and magnitude of the error using a given measurement method, and are therefore important for the proper design and interpretation of epidemiologic studies using that method. Currently, the epidemiologic area that is most actively engaged in the conduct of calibration studies is nutrition, probably because of the profound problems in measuring dietary intake. Most of the calibration studies that have been conducted in this area have been associated with a larger epidemiologic study using the same measurement instrument. In early

calibration studies it was assumed that one should attempt to measure the usual dietary intake of an individual as accurately as possible, to act as a gold standard comparison with the more approximate instrument to be used in the larger study (e.g., Willett et al., 1985). However, this wisdom was challenged when statisticians considered how such data would be used to aid interpretation of the main study. First, Carroll et al. (1984) and Rosner et al. (1989) demonstrated that a valid measurement error adjustment of the relative risk estimates from the main study can be made even when the calibration study does not include a very accurate measure of usual dietary intake. It would be sufficient to include a measure of intake that is simply *unbiased*. This meant that it was not imperative to use many repeat measurements of dietary intake so as to greatly increase the precision of the measure. Then, Kaaks, et al. (1995) and Stram, et al. (1995) considered the variance of the estimated relative risks adjusted for measurement error using such data from a calibration study. They discovered that if one has a choice between increasing the number of repeat measurements per individual or increasing the number of individuals, then, under assumptions of equal cost, the variance is minimized by maximizing the number of individuals in the calibration study. Stram, et al. (1995) also investigate the optimal strategy when costs are not equal.

The primary aim of a calibration study may not be exactly the same in each case. In this paper we consider four possibilities.

- (a) The aim that has been most frequently described is to use information from the calibration study to adjust the relative risks estimated from the main epidemiologic study for the measurement error associated with use of the FFQ (Kaaks, et al., 1995). It is well-known that measurement error biases the estimated relative risks and this motivates the need for adjustment.
- (b) Another possible aim is to estimate the sample size required in the main study. The required sample size depends heavily on the degree of measurement error associated with the FFQ (Freedman, Schatzkin and Wax, 1990), so there is good reason to check on this before proceeding with the main epidemiologic study. In this case it is important that the calibration study be conducted and evaluated before the main study proceeds.
- (c) A third possible aim is to estimate the correlation between FFQ intake and true usual intake. This could be of crucial interest if the FFQ has been modified extensively from previous versions or is to be used in a new population from which little previous data have been obtained. Very low correlations might persuade the investigators to postpone the main study, pending improvements in the design of the FFQ or in the way it is presented to study

participants.

- (d) A fourth possible aim is to estimate the slope of the regression of FFQ intake on the usual intake. This parameter is of importance in assessing the patterns of bias that might exist with use of the FFQ.

### **1.3 The American Association for Retired People (AARP) Study**

The National Cancer Institute and American Association for Retired People (AARP) are collaborating in conducting a large prospective nutritional epidemiological study, in which members of the AARP will report information on their dietary habits and will be followed to ascertain new diagnoses of cancer. The motivation for the study was: firstly, the degree of disagreement and controversy over the results of previous epidemiologic studies of diet and cancer, particularly breast cancer (Prentice, et al., 1988); secondly, the limited range of intakes of major macronutrients, such as fats, in previously studied cohorts (Hebert and Miller, 1988); and thirdly, the need for large numbers of cancer cases to occur during follow-up for the detection of small but important observed relative risks. The last point is emphasized by noting that a true relative risk of 2.0 can typically be reduced by dietary measurement error to an observed relative risk of 1.25 (Freudenheim and Marshall, 1988).

The design of the AARP study involves a two-stage sampling. Firstly, a large number of randomly selected members are sent a FFQ to complete. Secondly, a group of the respondents are selected using stratified random sampling on the basis of their reported intake on a selected macronutrient of interest, e.g. fat. The stratified sampling ensures that subjects with extremely high or low reported intakes have a high probability of being selected, while those with reported intakes closer to the average would have a lower probability of selection. Using percent calories from fat as the intake measure and five strata of intake ( $< 25\%$ ,  $25\%-32.5\%$ ,  $32.5\%-40\%$ ,  $40\%-47.5\%$ ,  $> 47.5\%$ ), the initially estimated required sample size for the cohort is 350,000 (Freedman, et al., 1991).

The calibration study is an important part of this study particularly because there is not wide experience with the results of mailing dietary questionnaires. There are three main aims of the calibration study: to check on the correlation between the reported FFQ intake and true usual intake to see if the mailed responses to the questionnaire have adequate “validity”; to check on the estimated sample size required in the main study; and to correct relative risk estimates from the main study.

Two options for the design of the calibration study suggest themselves. First, one might simply

take a simple random sample of the same AARP population and ask them to complete one or more FFQ's and one or more FR's or alternatively, one might design the calibration study to parallel the main study and preferentially select those individuals who report extreme intakes on their FFQ.

#### 1.4 Questions Posed in this Paper

In this paper, we analyze two aspects of the design of the calibration study. First, do we gain or lose efficiency by taking a stratified random sample? Second, is it better to obtain many FR's from a moderate number of individuals or only a small number of FR's on a larger number of individuals? Many researchers take the first option, so that they can characterize usual intake for each individual as accurately as possible. However, Kaaks, et al. (1995) argues that to achieve optimal adjustment of relative risks it is better to take only 1 FR per person and thus maximize the number of persons in the calibration study. Also, Rosner and Willett (1988) show that for estimating the correlation between a FFQ and usual intake, the optimal design depends on the amount of error in the FR's. We investigate these design options to see whether the optimal strategy depends on the different possible aims of a calibration study, as described in section 1.2.

The second question concerns analysis. It is general folklore that the method of moments and normal-theory maximum likelihood estimates in linear measurement error models have approximately the same efficiencies, and hence the former is useful because it has explicit formulae. We investigate these methods for the case when we sample from strata defined by the values of the response, and as described below we show that the usual folklore is seriously in error.

The paper is organized as follows. In section 2, we discuss the linear measurement error model, and place the AARP calibration study into this framework. In section 3, we discuss estimation in the context of missing data. The main conceptual device is to place linear errors-in-variables estimation into the framework of unbiased estimating functions. Using results of Rotnitzky and Robins (1995), we also show how to obtain the asymptotically optimal estimator which makes no distributional assumption. Sections 4 and 5 contain numerical results. In section 6 we discuss the implications of our results. Some technical details are collected into the appendix.

## 2 STATISTICAL MODEL FOR CALIBRATION

As described previously, the AARP calibration study has two stages. At the first stage, for a large number  $M$  of individuals, nutrient intake is measured by a FFQ. In the substudy, on a smaller number  $n$  of individuals, the FFQ nutrient intake is *calibrated* against usual intake by measuring

nutrient intake with two or more food records or recalls (FR), possibly together with additional FFQ's.

Our analysis is based upon the general statistical calibration model of Freedman, Carroll and Wax (1991), which we now discuss. They allow for the possibility that one or more FFQ's are measured contemporaneously with FR's, and hence that the errors are correlated. Here we will assume the simpler case that the FFQ's and the FR's are measured sufficiently far apart that all errors are uncorrelated.

Consider persons randomly selected to participate in the calibration study. The individual reports diet using a FFQ on  $m_1$  occasions ( $m_1 \geq 1$ ) and using a FR on  $m_2$  occasions ( $m_2 \geq 2$ ). The model relating intake of some nutrient (e.g., % calories from fat) reported on FFQ's (denoted by  $Q$ ) and intake reported on FR's (denoted by  $F$ ) to long-term usual intake (denoted by  $T$ ) is a standard linear errors-in-variables model, namely

$$Q_j = \beta_0 + \beta_1 T + r + \epsilon_j; \quad j = 1, \dots, m_1; \quad (1)$$

$$F_j = T + U_j; \quad j = 1, \dots, m_2. \quad (2)$$

In model (1),  $r$  is called the *equation error* (Fuller, 1987). The terms  $\epsilon_j$  represent the within individual variation in FFQ's, while the  $U_j$  are the within individual variation in FR's.

For example, in the AARP calibration study, a FFQ will be obtained initially, and some months later a FR will be obtained, followed by a second FR obtained at least one month later. Then  $m_1 = 1$  and  $m_2 = 2$ . If, subsequently, a second FFQ is obtained, then  $m_1 = 2$ .

Among these random variables,  $T$  has mean  $\mu_t$  and variance  $\sigma_t^2$ ,  $U_j$  has mean zero and variance  $\sigma_u^2$ ,  $\epsilon_j$  has mean zero and variance  $\sigma_\epsilon^2$ , and  $r$  has mean zero and variance  $\sigma_r^2$ . Note the critical assumption that  $U_j$  has mean zero, i.e., that the FR provides an unbiased measurement of dietary intake. All random variables are uncorrelated, although the methods are easily extended to allow for correlation in the measurement errors  $\epsilon_j$  and  $U_j$  when a questionnaire is given nearly coincidental in time to a record or recall. The parameter  $\sigma_\epsilon^2$  cannot be estimated if  $m_1 = 1$ , i.e., if there are no replicated FFQ's, and in this case the remaining parameters are estimated by setting  $\sigma_\epsilon^2 = 0$ ; the estimate of  $\sigma_r^2$  then incorporates the contribution of  $\sigma_\epsilon^2$ . If  $m_2 = 1$ , then the measurement error variable  $\sigma_u^2$  cannot be estimated, and it well known that  $\beta_1$  cannot then be estimated (Fuller, 1987).

In what follows, it is convenient to reparametrize the problem in terms of means, variances and covariances. Make the definitions  $\theta_1 = E(Q)$ ,  $\theta_2 = E(T)$ ,  $\theta_3 = \text{var}(Q)$ ,  $\theta_4 = \text{cov}(Q, F)$ ,  $\theta_5 = \text{var}(F)$ ,  $\theta_6 = \text{cov}(F_1, F_2)$  and  $\theta_7 = \text{cov}(Q_1, Q_2)$ . Let  $\Theta = (\theta_1, \theta_2, \dots, \theta_7)^t$ , and let  $e_k$  be the

vector of  $k$  1's. All the model parameters can be obtained from  $\Theta$ , specifically

$$\begin{aligned}\beta_1 &= \theta_4/\theta_6; \quad \mu_t = \theta_2; \quad \beta_0 = \theta_1 - \theta_2\theta_4/\theta_6; \quad \sigma_t^2 = \theta_6; \\ \sigma_u^2 &= \theta_5 - \theta_6; \quad \sigma_r^2 = \theta_7 - \theta_4^2/\theta_6; \quad \sigma_\epsilon^2 = \theta_3 - \theta_7.\end{aligned}$$

If  $m_1 = 1$ , then  $\theta_7$  cannot be estimated and (using the convention that  $\sigma_\epsilon^2 = 0$ )  $\sigma_r^2 = \theta_3 - \theta_4^2/\theta_6$ .

The possible observed data are summarized as  $Z = (Q_1, \dots, Q_{m_1}, F_1, \dots, F_{m_2})^t$ , which has mean  $(\theta_1 e_{m_1}^t, \theta_2 e_{m_2}^t)^t$  and covariance matrix

$$\Sigma(\Theta) = \begin{bmatrix} \theta_3 I_{m_1} + \theta_7 (e_{m_1} e_{m_1}^t - I_{m_1}) & \theta_4 e_{m_1} e_{m_2}^t \\ \theta_4 e_{m_2} e_{m_1}^t & \theta_5 I_{m_2} + \theta_6 (e_{m_2} e_{m_2}^t - I_{m_2}) \end{bmatrix}. \quad (3)$$

### 3 THE TWO-STAGE STUDY AS A MISSING DATA PROBLEM

#### 3.1 Introduction

Because the AARP main study will preferentially select individuals who report more extreme levels of dietary intake, we may wish the calibration study to have similar composition. We therefore consider a calibration design where sampling is done in two-stages. At the first stage, we observe the FFQ's  $Q_{i1}$  for  $M$  individuals,  $i = 1, \dots, M$ . Then at the second stage, the *calibration study*, with probability  $\pi(Q_{i1})$  we observe the  $m_2$  FR's  $(F_{i1}, \dots, F_{im_2})$  and the remaining  $m_1 - 1$  FFQ's  $(Q_{i2}, \dots, Q_{im_1})$ . If an individual is selected into the calibration study we set  $\Delta_i = 1$ , and otherwise we set  $\Delta_i = 0$ . The sampling weights are  $w_i = 1/\pi(Q_{i1})$ , the inverses of the probabilities of selection. In typical applications, the size of the calibration study is fixed, say to  $n$  observations, so that the  $\Delta$ 's are correlated.

This formulation allows for simple random sampling by setting  $\pi(Q)$  to be a constant, i.e., independent of the report from the first FFQ. The classical linear measurement error model assumes complete sampling, so that all individuals participate in the calibration study, and hence  $\Delta_i = \pi(Q_{i1}) = 1$ .

It is important to observe that this formulation is that of a missing data problem, wherein the FR's and the supplementary FFQ's are missing for many individuals. As a result of the design, the data are missing at random, i.e., missingness depends only on the value of the first FFQ and not on the unobserved FFQ's or FR's.

The purpose of this section is to discuss various estimation strategies. In section 3.2 we discuss two basic estimating functions for complete data, one based on the method of moments and one based on maximum likelihood estimation. Section 3.3 describes adaptations of these estimating

functions which allow for the missing data pattern of the AARP study. Section 3.4 gives some explicit details of the AARP study, which form the basis of all our later calculations. In practice, the sampling probabilities and hence the sampling weights are unknown and must be estimated, see also Section 3.4.

### 3.2 Method of Moments and Model Robustness

The typical measurement error model formulation has no missing data. The problem then is the classical linear measurement error model covered so admirably by Fuller (1987). With no missing data, there are two types of estimates in common use:

- (a) Method of moments estimators expressed in terms of the model parameters  $\mu_x, \sigma_x^2, \dots$ , and thus indirectly in terms of  $\Theta$ . This is effectively the method used by Fuller (1987, pp 106–108), and we will take it to be the default measurement error analysis.
- (b) Maximum likelihood estimators assuming that all random variables are normally distributed, and expressed in terms of the model parameters  $\Theta$ .

With no missing data, these estimators can be expressed in terms of solutions to unbiased estimating equations. These methods solve equations of the form

$$0 = \sum_i \Psi(Z_i, \Theta). \quad (4)$$

In what follows,  $i$  refers to the individual,  $Q_{ij}$  is the  $j$ th FFQ for the  $i$ th individual,  $\bar{Q}_i$  is the within-individual mean, and similarly for  $F_{ij}$  and  $\bar{F}_i$ . Also,  $m_1$  is the number of FFQ's for each individual, and  $m_2$  is the number of FR's. We use the term  $I(m_1 > 1)$  to be the indicator that  $m_1 > 1$ .

With no missing data, the estimating function for the method of moments is given by

$$\Psi_{\text{mom}}(Z_i, \Theta) = \begin{bmatrix} \bar{F}_i - \mu_t \\ \sum_{j=1}^{m_2} (F_{ij} - \bar{F}_i)^2 - (m_2 - 1)\sigma_u^2 \\ \begin{pmatrix} 1 & \bar{F}_i \\ \bar{F}_i & \bar{F}_i^2 - \sigma_u^2/m_2 \end{pmatrix}^{-1} \begin{pmatrix} \bar{Q}_i \\ \bar{Q}_i \bar{F}_i \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\ (\bar{F}_i - \mu_t)^2 - \sigma_u^2/m_2 - \sigma_t^2 \\ (\bar{Q}_i - \beta_0 - \beta_1 \bar{F}_i)^2 - \sigma_\epsilon^2/m_1 - \beta_1^2 \sigma_u^2/m_2 - \sigma_r^2 \\ \sum_{j=1}^{m_1} (Q_{ij} - \bar{Q}_i)^2 - (m_1 - 1)\sigma_\epsilon^2 \end{bmatrix}.$$

When there is only one FFQ ( $m_1 = 1$ ), we set  $\sigma_\epsilon^2 = 0$  and remove the last component of this estimating function.

The estimating function for the maximum likelihood estimator when there are no missing data is given in the appendix, section 7.1.

### 3.3 Model Robustness and Missing Data

When there are no missing data, by solving (4) method of moments and the maximum likelihood estimators are consistent and asymptotically normally distributed without restrictions as to the distributions of the random variables in the model. Of course, the asymptotic distributions depend on the underlying random variables, so that normal-theory information standard errors are valid only if all random variables are normally distributed. Otherwise, the simplest technique is to use sandwich standard errors: this is an old idea dating back at least to Huber (1967). There is also a sandwich-type theory of likelihood-ratio tests, see Huber (1967) and Kent (1982).

With missing data, if the probability of selection into the calibration study is  $\pi(Q_{i1})$ , then we can construct consistent estimates as follows. For the method of moments, we use only the validation data and weight it inversely with the selection probabilities, thus solving

$$0 = \sum_{i=1}^M \Delta_i \Psi_{\text{mom}}(Z_i, \Theta) / \pi(Q_{i1}). \quad (5)$$

This approach is of course the well-known Horvitz-Thompson method (Horvitz and Thompson, 1952).

With missing data, the moments estimate obtained through solving (5) is still consistent and asymptotically normal even without assuming normality. The normal-theory maximum likelihood estimator, however, does not share this property. Here is a subtle point. This likelihood estimator, which ignores the missing data mechanism, may give inconsistent parameter estimates if the random variables are not all normally distributed. a brief explanation is given in the appendix, section 7.2.

One can modify the likelihood estimator using the Horvitz-Thompson device to make it distribution-free, just as in (5). However, an asymptotically more efficient distribution-free estimator can be derived as follows. The problem of estimating  $\Theta$  without making any distributional assumptions is *semiparametric*, in the sense that parametric restrictions are made on the relationship of the means and variances of what we have called  $Z$ , while the underlying distributions are nonparametric. Optimal estimation of  $\Theta$  in such a context has been discussed by Rotnitzky and Robins (1995). Here we discuss their methods and adapt them to our problem. We do not justify any of the theoretical claims made here, as they are either proved by Rotnitzky and Robins or simple consequences of their arguments.

Let  $R$  be the vector of all individual elements of  $Z$  and their cross products. For example, if  $m_1 = 1$  and  $m_2 = 2$ ,  $Z = (Q_1, F_1, F_2)^t$  and  $R = (Q_1, F_1, F_2, Q_1^2, Q_1 F_1, Q_1 F_2, F_1^2, F_2^2, F_1 F_2)^t$ . Let  $g(\Theta) = E(R)$  and  $\nu(\Theta) = R - g(\Theta)$ . Of course, since we have specified the mean and covariance

matrix of  $Z$ ,  $g(\Theta)$  can be computed without reference to underlying distributions. Define  $\Delta = 1$  if an observation is selected into the calibration study and  $\Delta = 0$  otherwise. Superscript “t” refers to matrix transpose. Define

$$\begin{aligned} L &= \left\{ \frac{\partial}{\partial \Theta^t} g(\Theta) \right\}^t; \\ \chi(Z, \Delta, \Theta) &= \frac{\Delta \{R - E(R|Q_1)\}}{\pi(Q_1)} + E(R|Q_1) - g(\Theta); \\ \mathcal{T}_1(\Theta) &= \text{cov} \{ \chi(Z, \Delta, \Theta) \}; \\ \mathcal{T}_2(\Theta) &= \left( L \mathcal{T}_1^{-1} L^t \right)^{-1}. \end{aligned}$$

Rotnitzky and Robins prove that the best that any semiparametric estimator of  $\Theta$  can achieve is an asymptotic covariance matrix of  $M^{-1} \mathcal{T}_2(\Theta)$ . They further show that the optimal estimating function for achieving this covariance matrix is to solve

$$0 = \sum_{i=1}^M L \{ \mathcal{T}_1(\Theta) \}^{-1} \chi(Z_i, \Delta_i, \Theta). \quad (6)$$

This development has one unfortunate catch, namely that, as defined, implementation of (6) is impossible because  $\mathcal{T}_1$  and even  $\chi$  itself depend upon the underlying distributions. There are effectively two ways to implement the procedure:

- (a) Use nonparametric regression techniques to estimate  $\mathcal{T}_1$  and  $\chi$ . This gives global *asymptotic* efficiency, but is computationally burdensome and it is unclear if the asymptotics will agree with small sample behavior.
- (b) Assume a parametric model for  $Z$  only for the purposes of calculating  $\mathcal{T}_1$  and  $\chi$ . The resulting estimate is (locally) efficient if the parametric model actually holds, and it can be shown that the estimate is consistent even if the assumed parametric model is not correctly specified.

Since for our purposes we are contrasting the various estimators at the normal distribution anyway, we have followed method (b) with  $Z$  assumed to have the multivariate normal distribution. In this case,  $\mathcal{T}_1$  and  $\chi$  have closed-form expressions which are easily calculated.

We have calculated the asymptotic covariance matrix of the optimal semiparametric estimator when  $Z$  is normally distributed, and found that in a wide variety of cases it is essentially the same as that obtained from the Horvitz-Thompson method of moments estimators. This may not be the case away from the normal distribution, and it is an interesting problem for further study to see if major differences arise with departures from the normal distribution.

### 3.4 The AARP Study, Missing Data and Sampling Weights

At the first stage of the AARP calibration study, FFQ's are mailed to several tens of thousands of members of the AARP, randomly selected within certain states. From these FFQ's, individuals reporting extreme "patterns of food intake" are preferentially selected into the calibration study. Suppose that the pattern of intake is quantified by the percent of energy intake contributed by fat (% Calories from Fat). The reported intakes from the FFQ will help to characterize the distribution of a single FFQ report on % Calories from Fat. Suppose we wish to include in the calibration study the following proportions of individuals: 20% having  $Q_{i1} \leq 25$ , 15% with  $25 < Q_{i1} \leq 32.5$ , 10% with  $32.5 < Q_{i1} \leq 40$ , 15% with  $40 < Q_{i1} \leq 47.5$ , and 40% with  $47.5 < Q_{i1}$ .

To get some idea of the sampling fractions needed to achieve this, we use the distribution of % Calories from Fat as estimated from the FFQ report in the 1987 NHIS and in the Women's Health Trial Vanguard Study (Henderson, et al., 1990). The estimated mean and standard deviation are 38.25 and 57.76, respectively, and the distribution appears to be reasonably close to normality. We then estimate (using the normality assumption) that in the AARP study the selection probabilities should be 1.0, .178, .064, .126 and .962 depending on whether the observed % Calories from fat lies in 0–25, 25–32.5, 32.5–40, 40–47.5 and  $> 47.5$ , respectively.

In our numerical work, we will consider the two cases depending on whether the sampling weights are known or estimated. Based on the description in the previous paragraph, to know the weights we require that the distribution of FFQ's is known. This is typically unreasonable in practice, but in the AARP study the initial sample size will be so large that at least at a first level of approximation the distribution of intakes from FFQ's will be effectively known, as well as the mean  $\theta_1$  and variance  $\theta_3$ .

In other studies, the initial survey of FFQ's will not be so large, and then  $\theta_1$ ,  $\theta_3$  and the sampling probabilities must be estimated for all but the normal-theory maximum likelihood estimate. This need not be a bad thing, because Robins, Rotnitzky and Zhao (1994) have shown that such estimation can improve the large-sample properties of Horvitz-Thompson estimators. The obvious nonparametric estimate of the sampling probability in each stratum formed by the initial FFQ's is the proportion of individuals in the stratum who are selected into the calibration study.

## 4 MORE INDIVIDUALS OR MORE FOOD REPORTS?

When designing the calibration study, there are options regarding the numbers of individuals and how many FR's each case completes. For example, one might obtain just 2 FR's on many individ-

uals, or obtain many FR's but on fewer individuals. Put simply, if one can afford to obtain 4,000 FR's, which is the better design option:

- (a) 4 FR's on each of 1,000 individuals, or
- (b) 2 FR's on each of 2,000 individuals?

Of course, these two designs are not strictly comparable in terms of cost, but they may be nearly so. A design such as (b) incurs more recruitment costs. However, the design (a) risks a high dropout rate due to study participants becoming progressively less cooperative; such dropouts not only may bias the analysis but lead to greatly increased costs by attempting to obtain complete records on each individuals. Another potential difficulty with observing many FR's on individuals is the possibility for systematic time trends.

As mentioned in the introduction, the parameters of direct interest in the AARP study are  $\rho_{QT} = \theta_4 / (\theta_3 \theta_6)^{1/2}$ , the correlation between intakes from a single FFQ and true usual intake, and the total number  $N$  of cancer cases that need to be observed in the main study to achieve 90% power for detecting a plausible and worthwhile effect. Also of interest is the slope  $\beta_1$ . Freedman, et al. (1990) give a formula for  $N$ .

Whether design option (a) or (b) is preferred may depend on the parameter being estimated as well as on the within individual error variance in FR's ( $\sigma_u^2$ ) relative to the sum of the variance of FFQ's about the line ( $\sigma_r^2$ ) plus the within individual error variance in FFQ's ( $\sigma_e^2$ ). If FR's are relatively precise, then it may be better to select option (b) and maximize the number of individuals in the calibration study. As the FR's become relatively less precise, a switch may occur and it may become more efficient to select design option (a) and take more replicates per person. The switch point may vary according to the parameter being estimated.

We investigate these points first theoretically and then via computer simulation. All calculations use parameters in the model (1)–(2) as estimated via the techniques of Freedman, Carroll and Wax (1991) for % Calories from Fat as determined by the 1987 NHIS and the Women's Health Trial Vanguard Study (WHTVS), namely  $\mu_t = 38.25$ ,  $\sigma_t^2 = 24.45$ ,  $\sigma_e^2 + \sigma_r^2 = 40.92$ ,  $\sigma_u^2 = 30.36$ ,  $\beta_0 = 5.95$  and  $\beta_1 = 0.83$ . These values are consistent with  $\text{var}(Q) = 57.76$  mentioned earlier. The WHTVS used food records; if 24-hour recalls are used,  $\sigma_u^2$  is typically larger, and to incorporate this we did calculations also in the case that  $\sigma_u^2 = 83.35$ , a number obtained by an analysis of the CSFII (Continuing Survey of Food Intake by Individuals) data from the U.S. Department of Agriculture.

## 4.1 Theoretical Calculations

We first consider theoretical calculations, which are based upon the classical technique of Fisher information theory for the maximum likelihood estimator (Cox and Hinkley, 1981). We assume that initially FFQ's are obtained on  $M$  randomly selected individuals, and then FR's are obtained on a calibration subsample of size  $n$ . The calculations are standard if this second stage is selected completely at random, and if  $M$  is infinite so as to essentially completely characterize the distribution of a single FFQ: we will use both assumptions in our theory. Computer simulations will be used to show that the same results apply even when selection into the calibration study depends on the initial FFQ, and even if  $M$  is finite.

The results of the theoretical calculations are displayed in Figure 1, where we compare design options (a) and (b), described in the previous subsection. We allowed the within individual error variance in FR's to vary between 0.0 and 150.0 (remember,  $\sigma_u^2 \approx 30.36$  for food diaries, while  $\sigma_u^2 \approx 83.35$  for 24-hour recalls). As a function of the measurement error variance  $\sigma_u^2$  in the FR's, this figure compares the ratio of the theoretical (asymptotic) standard deviation for estimates of three parameters of interest; the correlation  $\rho_{QT}$ , the slope  $\beta_1$  and the required number of cancer cases  $N$ , for  $n = 2000$  FR's and  $m_2 = 2$  replicates (option b) to  $n = 1000$  FR's and  $m_2 = 4$  replicates (option a). Values of this ratio which are greater than 1.0 indicate that it is better to obtain many FR's on fewer individuals.

The results in Figure 1 are instructive. For our estimate of the within individual variance of food diaries ( $\sigma_u^2 = 30.36$ ), we see that for estimating  $N$ ,  $\beta_1$  and  $\rho_{QT}$ , it is more efficient to obtain fewer food records on many individuals (the ratio is less than 1.0). However, for our estimate of the within individual variance of 24-hour recalls ( $\sigma_u^2 = 83.35$ ), we see that especially for  $\beta_1$ , it is more efficient to obtain more recalls on less individuals, and to a lesser extent the same holds for  $\rho_{QT}$ . Interestingly, and for the considered range of the within-individual variance of FR's, for determining the required number of cancer cases  $N$ , it is more efficient to obtain only 2 FR's, both for food diaries and for 24-hour recalls.

The calculations we have done are easily extended in principle to the maximum likelihood estimator under stratified random sampling. Using the theory of estimating equations (Huber, 1967), similar calculations can be performed for the method of moments under either form of sampling.

## 4.2 Simulations for Simple Random Selection

The simulations we have done all agree qualitatively with the theoretical calculations, even when the simulations are applied to stratified sampling (unlike the theory presented here). Numerical results are given in the top half of Tables 1 and 2 for food diaries and in Table 3 for 24-hour recalls. We have listed mean squared errors and standard deviations. There is a small technical problem with listing mean squared errors, because Fuller (1987) shows that in fact they do not exist theoretically. However, in our particular simulations this issue was not a problem, because we checked the results against a more robust measure of variation, the median absolute deviation from the median, and found no real differences from the results reported here.

The estimators reported are the normal-theory maximum likelihood estimator and the method of moments estimator, namely solving (5); similar results with respect to design considerations were found for the other distribution-free estimators.

For food records ( $\sigma_u^2 = 30.36$ ), using a larger number of records per individual and fewer individuals is clearly less efficient than using a smaller number of records per individual and more individuals, whether for estimating the slope  $\beta_1$ , the correlation  $\rho_{QT}$  or the required number of cases  $N$ . For food recalls ( $\sigma_u^2 = 83.35$ ), we still see that it is more efficient to use fewer rather than more records per individual for estimating the required number of cases, but  $m = 4$  records per individual appears to be somewhat more efficient than  $m = 2$  records per individual for estimating the slope  $\beta_1$  and the correlation  $\rho_{QT}$ .

## 4.3 Effects of Stratified Sampling

In the AARP calibration study, stratified sampling appears attractive, so as to parallel the stratified nature of the main study. Here we study the statistical issue of stratified versus completely random sampling into the calibration substudy (Tables 1, 2 and 3)

The results are striking. Both asymptotic theory (not reported here) and simulations indicate that for the distribution-free methods, stratification causes a decrease in efficiency of estimation, particularly for the slope  $\beta_1$ , and for some cases for the correlation  $\rho_{QT}$ . There is also some decrease in efficiency for estimating the required number of cancer cases  $N$ , although the effect is not large.

Exactly the opposite results obtain for the normal-theory maximum likelihood estimator. Here we see that there is not much effect due to the design for estimating the slope  $\beta_1$ , but now the stratified design leads to noticeably smaller variability in the correlation  $\rho_{QT}$  and the required number of cancer cases  $N$ .

#### 4.4 Number of FFQ's

Tables 1-3 are simulations based on an infinite number ( $M$ ) of initial FFQ's. Obviously, in practice  $M$  will be finite, so we ran simulations with  $M = 15,000$  initial FFQ's, which is the target for the AARP study. The results are reported in Table 4, and should be compared to Table 1. There are essentially no differences between the tabulated value for  $M = \infty$  and  $M = 15,000$ .

### 5 PARAMETRIC OR SEMIPARAMETRIC ANALYSIS?

As stated previously, under simple random sampling into the calibration study, it is common folklore that the distribution-free estimates and the normal-theory maximum likelihood estimate behave similarly. Tables 1-3 report results for random selection for the maximum likelihood estimator and the method of moments estimator (5), and while there are some differences they are typically fairly minor, as expected. The semiparametric efficient estimator (6) is equivalent to the maximum likelihood estimator in this case.

It is when selection into the calibration study depends upon the initial response that we see major differences (Tables 1-3). While we report results only for the method of moments estimator (5), the semiparametric efficient distribution-free method (6) gave similar results. Both are vastly inferior to the normal-theory maximum likelihood estimator. For estimating the slope  $\beta_1$ , the correlation  $\rho_{QT}$ , or the required number of cancer cases  $N$ , the maximum likelihood estimator has less than 50% of the variance of the semiparametric methods.

### 6 DISCUSSION

A major point of our paper is that calibration studies may not always be designed simply to provide data to adjust relative risks in a larger study, although we agree that such adjustment is indeed an important aspect. For example, calibration studies should be used in the development of new measurement instruments to test whether the new measurement provides improvement over currently used methods. In this context, the correlation between the instrument and the true measurement (cf.  $\rho_{QT}$ ) would be of primary interest, and measures of bias (cf.  $\beta_0$  and  $\beta_1$ ) would also be important. As in the AARP study, calibration substudies may also be planned as part of the design phase in which design assumptions are checked. In this case, a central question is whether the designed sample size of the main study is justified. Our paper therefore addresses the design of calibration studies from a wider perspective than heretofore.

With regard to the question of the number of food reports per individual in the calibration study, our results are summarized in Table 5. They suggest that the conclusions of Kaaks, et al. (1995) and Stram, et al. (1995) that fewer repeat measurements on more individuals provides greater efficiency, is not completely general. Indeed, this was already demonstrated by Rosner and Willett (1988), who showed that for estimating the correlation between a FFQ and usual intake, the size of the measurement error in the FFQ and the FR's determine the optimal strategy. We have demonstrated as well that the optimal balance of repeats and individuals depends on the primary aim of the calibration study, as well as on the within individual variation of the repeated measurements. We should note, however, that for the particular parameters of our simulations, there were no cases where the choice of two repeats per individual was much worse than four repeats on half the number of individuals, indicating that in our case the amount of within individual variation, even in 24-hour recalls, was not enough to depart from the policy of 'maximizing the number of individuals'. From Figure 1, though, it is clearly quite possible that such a policy could be seriously in error in other circumstances.

The results regarding the advisability of stratified sampling into calibration studies do not provide a clear answer, since gains in efficiency are made under one analysis strategy (maximum likelihood) and losses are made under the other strategy (method of moments). As we have emphasized, the likelihood approach is valid only if the parametric structure is correctly specified. Likelihood methods require statistical models for the distribution of the "true" variate  $T$ . There has traditionally been considerable concern in the measurement error literature about the robustness of estimation and inferences based upon parametric models for unobservable variates. Fuller (1987, page 263) discusses this issue briefly in the classical nonlinear regression problem, and basically concludes that the results of parametric modeling "may depend heavily on the (assumed) form of the ( $T$ ) distribution". In probit regression, Carroll, Spiegelman, Lan, Bailey, and Abbott (1984) report that if one assumes that  $T$  is normally distributed, and it really follows a chisquared distribution with one degree of freedom, then the effect on the likelihood estimate is "markedly negative". Similar results are reported by Schafer (1987). Essentially all research workers in the measurement error field come to a common conclusion: likelihood methods can be of considerable value, but the possible nonrobustness of inference due to model misspecification is a vexing and difficult problem.

The issue of model robustness is hardly limited to measurement error modeling. Indeed, it pervades statistics, and has led to the rise of a variety of semiparametric and nonparametric tech-

niques. There is simply no agreement in the statistical literature as to whether semi/nonparametric or parametric modeling is more appropriate. Many researchers strongly believe that one should make as few model assumptions as possible. The argument here is that any extra efficiency gained by parametric modeling is more than offset by the need to perform careful and often time-consuming sensitivity analyses. Other researchers believe that appropriate statistical analysis requires one to do one's best to model every feature of the data, arguing in our context that it makes little sense to needlessly double the variance of parameter estimates.

The obvious question is whether the maximum likelihood estimate is actually sensitive in this context to model misspecification. We have run simulations with  $T$  having a scaled and translated negative exponential distribution, and found that the normal-theory maximum likelihood estimate of  $\beta_1$  is badly biased downwards, and this translates into a bias in the estimate of  $\rho_{QT}$ . For instance, for the parameters in Table 1,  $\beta_1 = 0.83$  and  $\rho_{QT} = 0.54$ , while in the simulations the averages are 0.62 and 0.49, respectively. While these biases are considerable, we note that based on simulations the 5%-level Anderson-Darling test for normality has power over 80% for detecting the nonnormality caused by the nonnormal distribution of  $T$  for as few as 2,000 FFQ's, and for larger sample sizes such as in the AARP study the power is nearly 100%. The point here is that while a misspecified likelihood analysis leads to badly biased estimates, in practice it is not impossible to detect the model misspecification, even with the large amounts of measurement error inherent in nutritional intake data.

A practical question is whether one can ever reasonably assume normality. With a stratified design, of course, the observed FR's will not be normally distributed anyway, and so distributional modeling is easiest for the FFQ's. It is often the case that nutrition data are transformed directly to normality (Nusser, et al., 1995 give one such approach), and the analysis is then done on the transformed scale. If one is willing to assume that when transformed FFQ's are normally distributed so too are their (transformed) component parts  $T$  and  $\epsilon$  as well as the FR's, then the modeling issue is "solved".

For % Calories from Fat, the nutrient intakes from many data sets appear reasonably normally distributed. In Table 6 we review the evidence of five studies with various instruments, noting that for the eight situations surveyed, six are reasonably normally distributed (and pass the Anderson-Darling test with level  $> 0.05$ ), one exhibits light-tailedness (Nurses Health Study, 4-day diaries) and another appears to be heavy-tailed (NHANES, 24-hour recalls). The latter is the only situation where one might expect that a parametric analysis assuming normality might be badly biased.

Since, as we have argued above, percent Calories from Fat often do appear to follow a normal distribution, our results indicate that in our case it would make sense to adopt a maximum likelihood approach, and consequently stratified sampling would appear beneficial. In general, however, we are not foolhardy enough to recommend one or the other approach. The important point is that we have identified a practical problem in which there is a surprisingly large difference between parametric and semiparametric modeling.

## REFERENCES

- Bingham, S. A. (1991). Limitations of the various methods for collecting dietary intake data. *Ann. Nutr. Metab.*, 35, 117–127.
- Carroll, R. J., Freedman, L. and Hartman, A. (1995). The use of semiquantitative food frequency questionnaires to estimate the distribution of usual intake. *American Journal of Epidemiology*, to appear.
- Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T. and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, 71, 19–26.
- Cox, D.R. and Hinkley, D.V. (1981). *Theoretical Statistics*. Academic Press, London.
- Freedman, L., Schatzkin, A. and Wax, Y. (1990). The impact of dietary measurement error on planning the sample size required in a cohort study. *American Journal of Epidemiology*, 132, 1185–1195.
- Freedman, L. S., Carroll, R. J. and Wax, Y. (1991). Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology*, 134, 510–520..
- Freedman L.S., Schatzkin A. and Wax Y. (1991). Re: The impact of dietary measurement error on planning sample size required in a cohort study. The authors reply. *American Journal of Epidemiology*, 134, 1472-1473.
- Freudenheim J.L. and Marshall J.R. (1988). The problem of profound mismeasurement and the power of epidemiologic studies of diet and cancer. *Nutrition and Cancer*, 11, 243-250.
- Fuller, W. A. (1987). *Measurement Error Models*. John Wiley and Sons, New York.
- Henderson, M. M., Kushi, L. H., Thompson, D. J., et al. (1990). Feasibility of a randomized trial of a low-fat diet for the prevention of breast cancer: dietary compliance in the Women's Health Trial Vanguard Study. *Prev. Med.*, 19, 115–133.
- Hebert J.R. and Miller D.R. (1988). Methodologic considerations for investigating the diet-cancer link. *American Journal of Clinical Nutrition*, 47, 1068-1077.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium*, 1, 221–233.

- Kaaks R., Riboli E. and van Staveren W. (1995). Sample size requirements for calibration studies of dietary intake measurements in prospective cohort investigations. *American Journal of Epidemiology*, 142, 557-565.
- Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69, 19–27.
- Nusser, S. M., Carriquiry, A. L., Dodd, K. W. & Fuller, W. A. (1995). A semiparametric transformation approach to estimating usual intake distributions. *Journal of the American Statistical Association*, to appear.
- Prentice R.L., Kakar F., Hursting S., Sheppard L., Klein R. and Kushi L.H. (1988). Aspects of the rationale for the Women’s Health Trial. *Journal of the National Cancer Institute*, 80, 802-814.
- Rotnitzky, A. and Robins, J. M. (1995). Semiparametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics*.
- Rosner, B. and Willett, W. C. (1988). Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. *American Journal of Epidemiology*, 127, 377–386.
- Rosner B.A., Willett W.C. and Spiegelman D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within- person measurement error. *Statistics in Medicine*, 8, 1051-1070.
- Schafer, D. (1987). Covariate measurement error in generalized linear models. *Biometrika*, 74, 385-391.
- Stram D.O., Longnecker M.P. and Shames L. (1995). Cost-efficient design of a diet validation study. *American Journal of Epidemiology*, 142, 353-362.
- Willett W.C., Sampson L. and Stampfer M.J. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology*, 122, 51-65,

## 7 APPENDIX

### 7.1 Estimating Function for the Normal-Theory Maximum Likelihood

We first write the estimating function and its derivatives for the case that there are no missing data. Recall that  $\Sigma(\Theta)$  is the covariance matrix of all the data, see (3). Let the partial derivatives of this matrix with respect to an arbitrary  $\theta_j$  be given by

$$P_j = \frac{\partial}{\partial \theta_j} \Sigma(\Theta),$$

where the derivative is component-wise. Then, using matrix derivatives, the estimating function for the maximum likelihood estimator when computed on all the data  $Z$  can be shown to equal (the numerical ordering appears slightly odd but is convenient later)

$$\Psi_{\text{ml}}(Z, \Theta) = \begin{bmatrix} \Psi_{\text{ml},1}(Z, \Theta) \\ \Psi_{\text{ml},2}(Z, \Theta) \end{bmatrix} = \{\ell_1(Z, \Theta), \ell_3(Z, \Theta), \ell_2(Z, \Theta), \ell_4(Z, \Theta), \dots\}^t,$$

where

$$\ell_1(Z, \Theta) = \begin{pmatrix} e_{m_1} \\ 0 \cdot e_{m_2} \end{pmatrix}^t \Sigma^{-1}(\Theta) \left\{ Z - \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix} \right\};$$

$$\ell_2(Z, \Theta) = \begin{pmatrix} 0 \cdot e_{m_1} \\ e_{m_2} \end{pmatrix}^t \Sigma^{-1}(\Theta) \left\{ Z - \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix} \right\};$$

and for  $j \geq 3$ ,

$$\begin{aligned} \ell_j(Z, \Theta) &= -(1/2) \text{trace} \left\{ \Sigma^{-1}(\Theta) P_j \right\} \\ &+ (1/2) \left\{ Z - \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix} \right\}^t \Sigma^{-1}(\Theta) P_j \Sigma^{-1}(\Theta) \left\{ Z - \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix} \right\}. \end{aligned}$$

Again using matrix derivatives, the Hessian of the  $m_1$  FFQ's and the  $m_2$  FR's can be computed explicitly as follows. Let

$$A(\theta_1, \theta_2) = ZZ^t - Z \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix}^t - \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix} Z^t + \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix} \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix}^t.$$

The Hessian of  $(\ell_1, \ell_2, \dots, \ell_7)$  is a  $7 \times 7$  matrix with elements  $h_{jk}(\Theta)$ , where

$$\begin{aligned} h_{11} &= - \begin{pmatrix} e_{m_1} \\ 0 \cdot e_{m_2} \end{pmatrix}^t \Sigma^{-1}(\Theta) \begin{pmatrix} e_{m_1} \\ 0 \cdot e_{m_2} \end{pmatrix} \\ h_{12} &= - \begin{pmatrix} e_{m_1} \\ 0 \cdot e_{m_2} \end{pmatrix}^t \Sigma^{-1}(\Theta) \begin{pmatrix} 0 \cdot e_{m_1} \\ e_{m_2} \end{pmatrix} \\ h_{22} &= - \begin{pmatrix} 0 \cdot e_{m_1} \\ e_{m_2} \end{pmatrix}^t \Sigma^{-1}(\Theta) \begin{pmatrix} 0 \cdot e_{m_1} \\ e_{m_2} \end{pmatrix} \end{aligned}$$

and for  $j, k \geq 3$ ,

$$\begin{aligned} h_{1j} &= - \begin{pmatrix} e_{m_1} \\ 0 \cdot e_{m_2} \end{pmatrix}^t \Sigma^{-1}(\Theta) P_j \Sigma^{-1}(\Theta) \left\{ Z - \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix} \right\}; \\ h_{2j} &= - \begin{pmatrix} 0 \cdot e_{m_1} \\ e_{m_2} \end{pmatrix}^t \Sigma^{-1}(\Theta) P_j \Sigma^{-1}(\Theta) \left\{ Z - \begin{pmatrix} \theta_1 e_{m_1} \\ \theta_2 e_{m_2} \end{pmatrix} \right\}; \\ h_{jk} &= (1/2) \text{trace} \left\{ \Sigma^{-1}(\Theta) P_k \Sigma^{-1}(\Theta) P_j \right\} \\ &- (1/2) \text{trace} \left[ \Sigma^{-1}(\Theta) \left\{ P_k \Sigma^{-1}(\Theta) P_j + P_j \Sigma^{-1}(\Theta) P_k \right\} \Sigma^{-1}(\Theta) A(\theta_1, \theta_2) \right]. \end{aligned}$$

Now we consider the possibility of missing data. From Little and Rubin (1987), the maximum likelihood estimator does not take into account the selection probabilities, and hence solves

$$0 = \sum_{i=1}^M \left( \Delta_i \begin{bmatrix} \Psi_{\text{ml},1}(Z_i, \Theta) \\ \Psi_{\text{ml},2}(Z_i, \Theta) \end{bmatrix} + (1 - \Delta_i) \begin{bmatrix} \Psi_{\text{ml},3}(Q_{i1}, \Theta) \\ 0 \end{bmatrix} \right),$$

where

$$\Psi_{\text{ml},3}(Q_{i1}, \Theta) = \begin{bmatrix} (Q_{i1} - \theta_1)/\theta_3 \\ (2\theta_3^2)^{-1} \{(Q_{i1} - \theta_1)^2 - \theta_3\} \end{bmatrix}.$$

## 7.2 Inconsistency of the MLE for Nonnormal Distributions

Showing this fact algebraically is a somewhat unpleasant task in general, but a simple special case illustrates the main idea. Suppose that the mean and variance of  $Q$  are known; this never happens

exactly, but in the AARP study  $n > 300,000$ , and so for all realistic purposes the mean and variance of  $Q$  really is known. For simplicity, suppose that  $\sigma_u^2$  is known, and that  $m_2 = 1$ , i.e., there is only one FR. Recalling that  $\theta_1 = E(Q) = \beta_0 + \beta_1\mu_t$ ,  $\theta_2 = E(F) = \mu_t$ ,  $\theta_3 = V(Q) = \beta_1^2\sigma_t^2 + \sigma_r^2$ ,  $\theta_4 = \beta_1\sigma_t^2$  and  $\theta_5 = \sigma_t^2 + \sigma_u^2$ , the unknown parameters for a normal-theory likelihood analysis are  $(\theta_2, \theta_4, \theta_5)$ . By detailed algebra, it may be shown that the normal-theory maximum likelihood estimator of  $\theta_2$  must satisfy

$$0 = \sum_{i=1}^n \Delta_i \{F_{i1} - \theta_2 - (\theta_4/\theta_3)(Q_{i1} - \theta_1)\}.$$

By the usual theory of estimating equations, if all parameters can be estimated consistently then

$$0 = E[\Delta \{F - \theta_2 - (\theta_4/\theta_3)(Q - \theta_1)\}].$$

Remembering that  $E(\Delta|F, Q) = E(\Delta|Q) = \pi(Q)$  because the data are missing at random, we see that consistency requires that

$$0 = E[\pi(Q) \{E(F|Q) - \theta_2 - (\theta_4/\theta_3)(Q - \theta_1)\}]. \quad (7)$$

Note that (7) holds if  $\pi(Q) \equiv \pi$ , a constant. In general, however, because the function  $\pi(\cdot)$  is arbitrary, for (7) to hold we require that the regression of  $F$  on  $Q$  be linear. This reflects the distributional assumption of normality, and need not hold otherwise.

		Selection at Random, $\sigma_u^2 = 30.36$							
$n$	#FR's	Method of Moments				Maximum Likelihood			
		MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
200	8	0.1014	0.0562	2508	721	0.0851	0.0534	2509	722
400	4	0.0840	0.0431	2457	501	0.0706	0.0403	2461	503
800	2	0.0707	0.0358	2393	390	0.0648	0.0344	2396	390
500	8	0.0646	0.0359	2441	433	0.0541	0.0342	2442	433
1000	4	0.0497	0.0271	2397	302	0.0431	0.0260	2398	301
2000	2	0.0489	0.0239	2389	240	0.0443	0.0227	2391	240
800	8	0.0512	0.0279	2385	313	0.0408	0.0261	2386	312
1600	4	0.0407	0.0221	2390	240	0.0351	0.0211	2391	241
3200	2	0.0368	0.0181	2380	188	0.0337	0.0174	2380	188
		Selection on the Basis of FFQ, $\sigma_u^2 = 30.36$							
$n$	#FR's	Method of Moments				Maximum Likelihood			
		MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
200	8	0.1192	0.0581	2484	649	0.0785	0.0373	2415	403
400	4	0.0975	0.0443	2430	475	0.0656	0.0312	2395	322
800	2	0.0992	0.0430	2408	393	0.0673	0.0282	2385	249
500	8	0.0744	0.0357	2417	389	0.0482	0.0232	2387	250
1000	4	0.0629	0.0295	2398	302	0.0416	0.0193	2385	193
2000	2	0.0585	0.0255	2395	239	0.0389	0.0168	2384	157
800	8	0.0590	0.0288	2377	288	0.0373	0.0184	2380	193
1600	4	0.0496	0.0230	2389	221	0.0317	0.0148	2377	147
3200	2	0.0481	0.0209	2384	192	0.0311	0.0133	2375	125

Table 1: *Simulation results using Food Records (FR) and one Food Frequency Questionnaire (FFQ), with  $\alpha = 5.95$ ,  $\beta = 0.83$ ,  $\sigma_r^2 = 16.207$ ,  $\sigma_e^2 = 24.71$ ,  $\mu_t = 38.253$ ,  $\sigma_t^2 = 24.449$ ,  $\sigma_u^2 = 30.36$ . By “Selection on the Basis of FFQ” we mean stratified sampling within ranges of FFQ reported values. The value of  $n$  is the number of individuals in the calibration study. The terms  $\rho_{QT}$  and  $\hat{N}$  refer to the estimate of the correlation between an FFQ and usual intake and the estimated required number of cancer cases, respectively. The method of moments uses the standard parametrization as defined in the text.*

		Selection at Random, $\sigma_u^2 = 30.36$							
		Method of Moments				Maximum Likelihood			
$n$	#FR's	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
200	8	.0921	.0498	2499	624	.0733	.0451	2496	618
400	4	.0693	.0373	2436	448	.0610	.0349	2434	444
800	2	.0651	.0304	2403	333	.0615	.0290	2405	332
500	8	.0544	.0299	2391	355	.0444	.0269	2390	349
1000	4	.0442	.0230	2396	260	.0373	.0210	2395	258
2000	2	.0405	.0198	2393	209	.0373	.0186	2393	206
800	8	.0421	.0236	2391	273	.0340	.0214	2388	270
1600	4	.0362	.0193	2377	208	.0305	.0174	2375	203
3200	2	.0336	.0161	2380	160	.0309	.0151	2380	167
		Selection on the Basis of FFQ, $\sigma_u^2 = 30.36$							
		Method of Moments				Maximum Likelihood			
$n$	#FR's	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
200	8	.1109	.0537	2471	659	.0700	.0339	2413	392
400	4	.0919	.0419	2451	478	.0583	.0263	2396	281
800	2	.0948	.0392	2413	372	.0620	.0251	2386	225
500	8	.0682	.0331	2410	393	.0431	.0209	2395	239
1000	4	.0565	.0265	2387	289	.0357	.0167	2375	181
2000	2	.0572	.0244	2387	236	.0367	.0151	2376	144
800	8	.0549	.0271	2421	309	.0337	.0170	2389	193
1600	4	.0442	.0211	2373	226	.0293	.0136	2376	146
3200	2	.0453	.0200	2386	192	.0307	.0123	2373	112

Table 2: Simulation results using Food Records (FR) and two Food Frequency Questionnaires (FFQ), with  $\alpha = 5.95$ ,  $\beta = 0.83$ ,  $\sigma_r^2 = 16.207$ ,  $\sigma_c^2 = 24.71$ ,  $\mu_t = 38.253$ ,  $\sigma_t^2 = 24.449$ ,  $\sigma_u^2 = 30.36$ . See previous table for definition of terms.

		Single FFQ							
		Selection at Random, $\sigma_u^2 = 83.35$							
		Method of Moments				Maximum Likelihood			
$n$	#FR's	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
500	8	.0759	.0408	2404	473	.0679	.0394	2406	473
1000	4	.0738	.0363	2403	379	.0689	.0349	2407	378
2000	2	.0832	.0356	2387	330	.0817	.0354	2391	330
		Selection on the Basis of FFQ, $\sigma_u^2 = 83.35$							
		Method of Moments				Maximum Likelihood			
$n$	#FR's	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
500	8	.1022	.0443	2419	443	.0632	.0288	2396	291
1000	4	.1051	.0424	2402	358	.0662	.0272	2383	231
2000	2	.1392	.0517	2404	340	.0814	.0302	2372	206
		Two FFQ's							
		Selection at Random, $\sigma_u^2 = 83.35$							
		Method of Moments				Maximum Likelihood			
$n$	#FR's	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
500	8	.0657	.0342	2397	396	.0567	.0312	2395	390
1000	4	.0650	.0303	2401	326	.0628	.0296	2404	326
2000	2	.0816	.0324	2398	287	.0800	.0318	2390	286
		Selection on the Basis of FFQ, $\sigma_u^2 = 83.35$							
		Method of Moments				Maximum Likelihood			
$n$	#FR's	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
500	8	.0917	.0413	2410	451	.0594	.0263	2383	273
1000	4	.0912	.0386	2420	374	.0588	.0240	2395	227
2000	2	.1247	.0464	2420	359	.0801	.0284	2393	194

Table 3: *Simulation results using 24-hour Recalls (24-FR), with  $\alpha = 5.95$ ,  $\beta = 0.83$ ,  $\sigma_r^2 = 16.207$ ,  $\sigma_\epsilon^2 = 24.71$ ,  $\mu_t = 38.253$ ,  $\sigma_t^2 = 24.449$ ,  $\sigma_u^2 = 83.35$ . By “Selection on the Basis of FFQ” we mean stratified sampling within ranges of FFQ reported values. The value of  $n$  is the number of individuals in the calibration study. The terms  $\rho_{QT}$  and  $\hat{N}$  refer to the estimate of the correlation between an FFQ and usual intake and the estimated required number of cancer cases, respectively.*

		Single FFQ, 2 FR's							
		Selection on the Basis of 15,000 Initial FFQ's, $\sigma_u^2 = 30.36$							
		Method of Moments				Maximum Likelihood			
$n$	#FR's	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
500	8	.0744	.0368	2393	384	.0492	.0236	2381	250
1000	4	.0600	.0287	2390	299	.0405	.0191	2377	192
2000	2	.0602	.0261	2407	244	.0384	.0166	2381	158
		Single FFQ, 2 24-FR's							
		Selection on the Basis of 15,000 Initial FFQ's, $\sigma_u^2 = 83.35$							
		Method of Moments				Maximum Likelihood			
$n$	#FR's	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$	MSE $\beta$	MSE $\rho_{QT}$	mean $\hat{N}$	s.d. $\hat{N}$
500	8	.0966	.0443	2426	453	.0631	.0292	2391	292
1000	4	.0979	.0415	2444	394	.0659	.0271	2392	245
2000	2	.1288	.0484	2416	359	.0851	.0304	2389	207

Table 4: *Simulation results using two Food Records (FR,  $\sigma_u^2 = 30.36$ ) and two 24-hour Recalls (24-FR,  $\sigma_u^2 = 83.35$ ), with  $\alpha = 5.95$ ,  $\beta = 0.83$ ,  $\sigma_r^2 = 16.207$ ,  $\sigma_c^2 = 24.71$ ,  $\mu_t = 38.253$ ,  $\sigma_t^2 = 24.449$ . Here we use stratified sampling within ranges of FFQ reported values, where the initial survey consists of 15,000 FFQ's. The value of  $n$  is the number of individuals in the calibration study. The terms  $\rho_{QT}$  and  $\hat{N}$  refer to the estimate of the correlation between an FFQ and usual intake and the estimated required number of cancer cases, respectively.*

PROBLEM	Min #	Optimal #
Correcting relative risks using regression calibration	1	1
Estimating required number of cases to detect effect at a given power	1 or 2, depending on method used	Same as Min #
Estimating correlation $\rho_{QT}$	2	No uniform answer
Estimating slope $\beta_1$	2	No uniform answer

Table 5: For various problems, the minimum number of FR's required in a calibration study (Min #) and the optimal number (Optimal #).

s.e. n=2000, 2 FD/FR / s.e. n=1000, 4 FD/FR

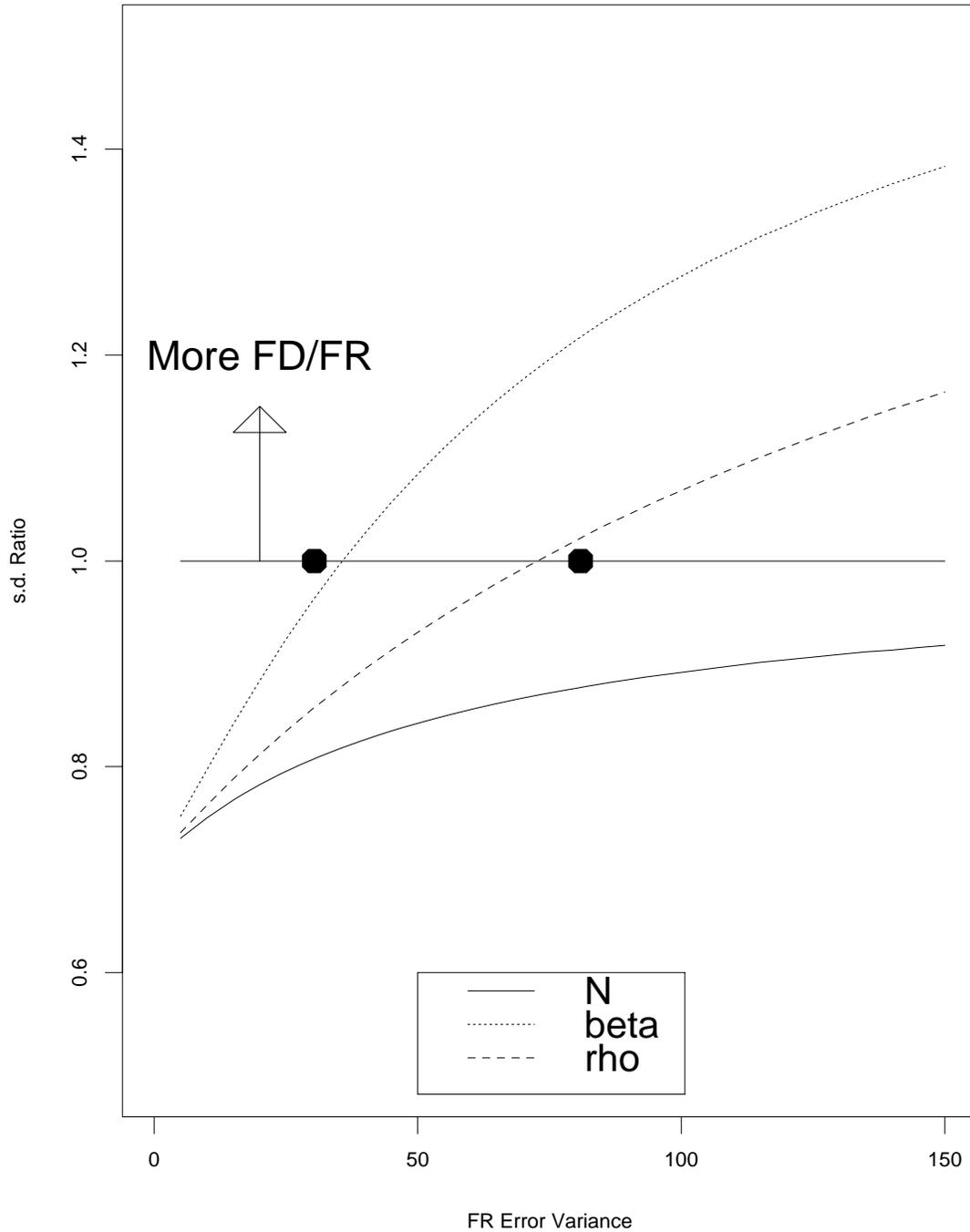


Figure 1: A two-stage calibration study, such as contained within the AARP, with a large number of initial FFQ's, under the parameter configurations  $\sigma_r^2 = 40.92$ ,  $\mu_t = 38.25$ ,  $\sigma_t^2 = 24.45$ ,  $\beta_0 = 5.95$  and  $\beta_1 = 0.83$ . As a function of the measurement error variance  $\sigma_u^2$  in the FR's, this figure compares the ratio of the asymptotic standard deviation for estimates of  $\rho_{QT}$  ("rho in the figure"),  $\beta_1$  ("beta") and the sample size  $N$  ("N"), for  $n = 2000$  FR's and  $m_2 = 2$  replicates to  $n = 1000$  FR's and  $m_2 = 4$  replicates.

Study	Instrument	# Instruments per participant	Sample Size	Skewness	Kurtosis	A-D Test
WISH	FFQ	1	271	-0.12	3.24	0.68
WISH	24-hour recall	6	271	-0.31	3.28	0.49
CSFII	24-hour recall	3	1705	0.01	3.42	0.71
NHANES	24-hour recall	1	3145	0.08	3.65	1.61
NHS	FFQ	1	168	0.13	3.60	0.29
NHS	4-day diary	4	168	-0.29	2.38	0.91
WHTVS	FFQ	3	86	0.16	2.58	0.42
WHTVS	4-day diary	2	86	-0.55	3.35	0.47

Table 6: *Tests for normality for the variable % Calories from Fat for various nutrition data sets. The 5% significance level for the A-D (Anderson Darling) test is 0.78. Definition of acronyms: WISH (Women’s Interview Survey of Health); CSFII (Continuing Survey of Food Intake by Individuals); NHANES (National Health and Nutrition Examination Survey); NHS (Nurses’ Health Study); WHTVS (Women’s Health Trial Vanguard Study).*

## NOTE TO READERS

Throughout the text, vectors and matrices have been underlined and boldfaced. The symbols in question are as follows:

$0$

$I$

$\Theta$

$\chi$

$\Psi$

$Z$

$R$

$g$

$\mathcal{T}$

$L$

$\nu$

$A$

$\Sigma$

$P$

$e$