# Estimating the Relationship Between Dietary Intake Obtained From a Food Frequency Questionnaire and True Average Intake

by

Laurence S. Freedman
National Cancer Institute

Raymond J. Carroll*
Texas A & M University

Yohanan Wax
Hebrew University

Technical Report # 90–35C

Department of Statistics
Purdue University

July, 1990

*The research of Professor Raymond J. Carroll is supported in part by the National Science Foundation Grants DMS–8923071 and DMS–8717799 at Purdue University.

△

Prepared for submission to *American Journal of Epidemiology* - 7/25/90

# ESTIMATING THE RELATIONSHIP BETWEEN DIETARY INTAKE OBTAINED FROM A FOOD FREQUENCY QUESTIONNAIRE AND TRUE AVERAGE INTAKE

Laurence S. Freedman,[1] Raymond J. Carroll,[2] AND Yohanan Wax[3]

[1] Biometry Branch, DCPC
National Cancer Institute
Bethesda, Maryland

[2] Department of Statistics
Texas A&M University
College Station, Texas

[3] Department of Statistics
Hebrew University
Jerusalem 91905, Israel

SEND REPRINT REQUESTS TO:

Mr. Laurence S. Freedman
Biometry Branch, DCPC
National Cancer Institute
Executive Plaza North, Suite 344
Bethesda, MD 20892

**ABSTRACT**
(not to exceed 200 words)

Knowledge of the regression relationship between dietary intake reported on a food frequency questionnaire and true average intake is useful in interpreting results from nutritional epidemiologic studies and in planning such studies. Studies which validate a questionnaire against a food record may be used to estimate this regression relationship provided the food record is completed by each subject on at least two occasions. Using data from the pilot study of the Women's Health Trial, we show how variation in diet over time and intraindividual correlation between a questionnaire and food record obtained close together in time affects the estimation of the regression. Our method provides estimates of the regression slope and the questionnaire "bias" that are corrected for these effects, together with standard errors. In an Appendix we provide a SAS program to carry out the analysis.

[137 words]

*Key words:* (not more than eight)

## INTRODUCTION

Food frequency questionnaires are now quite commonly used to obtain estimates of individuals' dietary intake and so relate diet to the development of various diseases. Unfortunately the data gathered from a questionnaire can give estimates of dietary intake which are seriously in error. Questionnaires therefore need to be validated against a more reliable source of information, such as a detailed record of all food eaten by an individual over a period of one day or more (1). It is quite common to report the correlation between a questionnaire and the food record. However, the food record may not be a good estimate of average intake over a long period, as the food record represents intake over a very short period and there is considerable variation in intake over time (2,3). Thus several food records may be taken and the questionnaire validated against the average of these, with the assumption that this average is precise enough to substitute for the true average intake.This assumption is often unjustified. Methods are available to correct for the imprecision (2,3). Salvini et al. (4) have recently employed these methods to correct correlations for week-to-week variation in food consumption.

A second statistical phenomenon which can affect the validation is the presence of correlation between intakes from questionnaires and food records taken at approximately the same time. For example if a questionnaire and food record are taken twice each, once at the beginning and once at the end of the a study, then the first questionnaire may agree more closely with the first food record than with the second one. Beaton et al. (3) term this phenomenon "intraindividual correlation" and show that it will lead to apparent (but not real) enhancement of the validity of the questionnaire. They give a method to adjust the estimates of correlation of questionnaire on food record for the presence of intraindividual correlation.

However, the method seems rarely to be used. Willett (5) remarks that the assumption of no intraindividual correlation "does not appear to be seriously violated in most instances."

While correlation is a useful measure of the validity of a questionnaire, it is often more informative to obtain a regression equation between quantitative and food record. Such regression relationships are helpful in examining the nature of possible biases in the questionnaire, for example the possible tendency for those with low-calorie diets to overestimate the amount they are eating. Moreover, knowledge of the regression relationship is useful in interpreting the published results of nutritional epidemiologic studies (6), planning epidemiological studies of diet and disease (7), and in correcting the results of such studies for dietary measurement error (8).

As with correlation coefficients, estimates of regression slopes and intercepts can be affected by variation in diet and intraindividual correlation. In this paper we demonstrate the importance of adjustments to regression slopes and intercepts, as well as correlations, for variation in diet and intraindividual correlation. We use data from the pilot phase of the Women's Health Trial (9), which include a food frequency questionnaire completed twice and a food record completed three times by each subject. Analysis of these data require a statistical model which incorporates the measurements from multiple questionnaires and food records, allowing estimation of regression parameters and their standard errors. In an Appendix we describe the full statistical model which underlies our analysis and provide a SAS program for performing the calculations.

## HOW VARIATION IN DIET AFFECTS VALIDATION OF THE QUESTIONNAIRE

Validation of a questionnaire requires a comparison of the intakes obtained from the questionnaire with the average intakes obtained from several food records. If correlations are used to assess the strength of this relationship, then the higher the correlation, the better is the validation. If a regression of questionnaire versus average food record is presented, then desirable characteristics of the regression would be an intercept of approximately zero and a slope of approximately 1. These would indicate that the questionnaire is "unbiased," i.e. on average the questionnaire intake will equal the average food intake. Even when these conditions are met the questionnaire intakes will typically be quite widely spread about the average food intake, with consequently low correlation between questionnaire and food record. However in many cases validation shows, besides low correlations (between 0.3 and 0.7), slopes of less than 1. This finding has been termed the "flattened slope" effect (10).

Different interpretations of the flattened slope effect exist. One interpretation is that the effect is due to reporting bias. People with low intakes tend to overestimate their intake on a questionnaire, whereas those with high intakes tend to underestimate. This explanation is attractive since it coincides with our perception that people with extreme eating habits would like to be thought to behave somewhat nearer to the norm. A second interpretation is that the effect is due to the "error" in the food record, i.e. variation in food intake over time (11). A third interpretation of the effect is that a scale difference exists in the conversion of questionnaire and food record items to estimated nutrient intakes.

In the following we introduce a simple statistical model which will help us to study quantitatively the influence of variation in food intake over time on the estimated regression slope. This model involves only a single application of the questionnaire for each subject. The full model described in the Appendix incorporates repeated applications of the questionnaire.

Suppose that for any subject a food record intake F is obtained at a time randomly chosen in a given interval during which average intake is to be estimated. Let the food record intake F be equal to the true average intake T plus some random error u. We assume that u is independent of T and its variance is the same for any value of T. These assumptions accord quite well with data on percent calories from fat intake and with the logarithmic transformation of intakes of nutrients such as calories and protein. We can show under these assumptions that if $\beta_{QT}$ is the slope relating intake reported on a single questionnaire Q to true average intake T then the regression slope $\beta_{QF}$ of questionnaire on food record is given by

$$\beta_{QF} = \beta_{QT} \cdot \frac{\text{var } T}{\text{var } (F)} = \beta_{QT} \left( \frac{\sigma^2_{BF}}{\sigma^2_{BF} + \sigma^2_{WF} / n_F} \right) \qquad (1)$$

where $\sigma^2_{BF}$ and $\sigma^2_{WF}$ are the between (inter) and within (intra) components of variance for the food record and $n_F$ is the number of food records included in the average. The multiplication factor $\sigma^2_{BF} / \left( \sigma^2_{BF} + \sigma^2_{WF} / n_F \right)$ is always less than or equal to 1, so equation (1) describes the tendency for the regression slope of questionnaire on food record to be less than the slope of questionnaire on truth. In particular if the latter is 1 then $\beta_{QF}$ will tend to be less than 1, as in the observed flattened slope effect.

From equation (1) we may obtain a formula for calculating the true regression slope $\beta_{QT}$ of questionnaire on average intake. Thus

$$\beta_{QT} = \beta_{QF}\left(1 + \frac{\sigma^2_{WF}}{n_F \cdot \sigma^2_{BF}}\right) \qquad (2)$$

It may be thought that since the value F comes from the average of several food records the multiplication factor in equation (2) is close enough to 1 to make little difference. However calculations with data gathered in the pilot phase of the Women's Health Trial indicate that this factor can be substantially greater than 1. In this study women volunteers, aged 45 to 69 years, were randomized to one of two groups: those counselled to eat a low fat diet and those who were not so counselled. Food records were satisfactorily completed by 85 subjects in the Usual Diet group at the beginning of the study, 6 months, 1 year and 2 years. Records completed by the subjects in the Low Fat Diet group were not included in this analysis. Since the reported average caloric intakes of the Usual Diet Group appeared at baseline to be higher than after entry to the study, the baseline record was excluded and the average of the remaining 3 food records was used. Thus $n_F = 3$. Estimates of the multiplication factor in equation (2) ranged from 1.20 to 1.35 for the 5 nutrients examined (Table 1). A semi-quantitative food frequency questionnaire developed by Block *et al* (12) was completed at 1 and 2 years. At 1 year the women were instructed to report intake over the previous 6 months; at 2 years they were instructed to report the previous 12 months intake. Relating these questionnaire responses separately to the average of the food record, the flattened slope effect ($\beta_{QF} < 1$) is evident for all 5 nutrients (Table 1). Estimates of the true slope $\beta_{QT}$ are also shown. These values are generally near 1 and indicate that the flattened slopes may be explained by variation in diet. However this conclusion is premature since

we have not yet accounted for intraindividual correlation. This is done in the next section.

Besides affecting regression slopes, variation in the food record over time also affects correlations. If $\rho_{QF}$ is the correlation between questionnaire and food record and $\rho_{QT}$ is the correlation between questionnaire and truth, then under the same assumptions as above:

$$\rho_{QF} = \rho_{QT} \sqrt{\frac{\sigma_{BF}^2}{\sigma_{BF}^2 + \sigma_{WF}^2 / n_F}} \qquad (3)$$

Thus $\rho_{QF}$ is modified by the square root of the multiplication factor pertaining to equation (1). Since the square root brings the factor closer to 1, $\rho_{QF}$ is not affected as strongly as $\beta_{QF}$ by variation in diet. Table 2 shows estimates of $\rho_{QF}$ and $\rho_{QT}$ for the 1 year and 2 year questionnaires. Confidence intervals for the estimate of $\rho_{QT}$ can be obtained by a method described by Rosner and Willett (13).

## INTRAINDIVIDUAL CORRELATIONS

In the discussion of the previous section an unstated assumption was that within an individual the questionnaire and the food record are uncorrelated. The intraindividual correlation between a questionnaire and food record which are administered in proximity is a measure of the tendency for the two to be closer in agreement than other questionnaires and food records administered at times more distant. If a study were designed so that each individual i (i = 1, ... ,n) were assessed by a questionnaire and food record at times j (j = 1, ... ,n) then the intraindividual correlation would be estimated by:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(Q_{ij}-\overline{Q}_i\right)\left(F_{ij}-\overline{F}_i\right)\Big/\sqrt{\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\left(Q_{ij}-\overline{Q}_i\right)^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\left(F_{ij}-\overline{F}_i\right)^2\right)}$$

where $Q_{ij}$, $F_{ij}$ are subject i's responses to questionnaire and food record at time j and $\overline{Q}_i$, $\overline{F}_i$ are the mean responses of the $i^{th}$ subject. Note that all variances and covariances are calculated within individual in this estimation. Contrary to the assumption of the previous section, intraindividual correlations due to proximity of a questionnaire and food record are not only non-zero, but, as shown below, can be substantial.

Taking two questionnaires or two food records in proximity, or taking a food record and a questionnaire in proximity is likely to induce correlated errors for two reasons. Firstly, subjects may remember their responses to the first inquiry and repeat them on the second inquiry. This will affect primarily questionnaire data. Secondly, since both are taken close together they will not capture the full variation of diet over time. They are therefore more likely to be in agreement with each other than with other questionnaires or food records taken at more distant time points. This will apply to both questionnaires and food record data.

For the data in our example the food records and questionnaires are taken at intervals of 6 months or 1 year. However the 1-year questionnaire and food record were taken in close proximity to each other, as were the 2-year questionnaire and food record. Table 3 displays the estimated intraindividual correlation between the questionnaire and its proximate food record. These estimates are obtained from a formula similar to that given above, but adjusted for the unequal number of questionnaires (2) and food records (3), as described in the Appendix. The correlation is negligible for carbohydrate, moderate for calories and protein, and high for saturated fat and percent calories from fat.

This correlation has an effect on the estimate of the regression slope of questionnaire on food record. Specifically if we denote the intraindividual correlation by r and the within component of variance for the questionnaire by $\sigma^2_{WQ}$ then we can write:

$$\beta_{QF} = \frac{\beta_{QT}\sigma^2_{BF} + r\,\sigma_{WF}\sigma_{WQ}/n_F}{\sigma^2_{BF} + \sigma^2_{WF}/n_F}, \qquad (4)$$

where $n_F$ is the number of food records taken.

Equation (4) is the same as equation (1) if r equals zero. Positive values of r tend to increase the observed slope of questionnaire on food record. From equation (4) we can obtain a formula for $\beta_{QT}$, namely:

$$\beta_{QT} = \beta_{QF}\left(\frac{1 + \sigma^2_{WF}}{n_F \cdot \sigma^2_{BF}}\right) - r\,\frac{\sigma_{WF}\,\sigma_{WQ}}{n_F \cdot \sigma^2_{BF}} \qquad (5)$$

From these results we see that if there is a positive intraindividual correlation, then $\beta_{QF}$ tends to be raised and thus the estimate of $\beta_{QT}$ is reduced. Thus positive intraindividual correlation has an effect in the opposite direction from variation of diet, which tends to reduce $\beta_{QF}$.

Intraindividual correlation also modifies the observed correlation between questionnaire and food record. This may be written as:

$$\rho_{QF} = \rho_{QT}\sqrt{\frac{\sigma^2_{BF}}{\sigma^2_{BF} + \sigma^2_{WF}/n_F}} + r\,\frac{\sigma_{WF}\,\sigma_{WQ}/n_F}{\sqrt{\left(\sigma^2_{BF} + \sigma^2_{WF}/n_F\right)\left(\sigma^2_{BQ} + \sigma^2_{WQ}\right)}} \qquad (6)$$

These results were applied to the Women's Health Trial data (Table 4). The estimates of $\beta_{QT}$ are, except one, all less than 1, some considerably so. This

reverses the earlier impression, obtained from assuming no intraindividual correlation, that the values of $\beta_{QT}$ were quite close to 1.

In Table 5 we show results from the statistical model (see Appendix) in which we combine the analysis of the first and second questionnaires to obtain overall estimates of $\beta_{QT}$ and $\rho_{QT}$. This method yields standard errors for $\beta_{QT}$. The estimates of $\beta_{QT}$ and $\rho_{QT}$ lie between the estimates for the individual questionnaires presented in Table 4.

## QUESTIONNAIRE BIAS

In the linear regression of questionnaire on true average intake the intercept would represent the average response to a questionnaire of a person truly eating nothing of the nutrient in question. This is not a particularly meaningful quantity and instead we report "bias." By "bias" we mean the difference between the estimated value of the questionnaire for a person truly eating the mean amount of nutrient in the population, and the mean itself.

We show estimates and standard errors of the "bias" (Table 5). These show that the questionnaire tends to underestimate intake of calories, protein, carbohydrates and saturated fat, the "biases" being negative and each more than twice its standard error. The estimated underestimation is around 0.1 on the log scale at the mean intake, equivalent to approximately 10% underestimation of the level of intake. The "bias" for percent calories from fat, however, is small and not significant.

## SUMMING UP

In this paper we report methodology for obtaining estimates of correlation, regression slope and "bias" of a food questionnaire with true average intake. For the five nutrients investigated we have found a "bias" in the questionnaire towards underestimation of intake, and a flattened slope effect (i.e. regression slope less than one). The bias is negligible for percent calories from fat but amounts to 10% underestimation for the other nutrients. The flattened slope effect will increase the percent underestimation for those consuming greater than average amounts of nutrient, but will tend to compensate for underestimation in those consuming less than average amounts.

Our analysis suggests that the flattened slope was not due to variation in diet in this set of data. Although the slope for an individual nutrient is not significantly less than 1, the combined set of "flattened slopes" suggests that the effect is real. A formal significance test could be devised for this hypothesis but would need to account for the correlation between the estimated slopes. Whether the flattened slope effect represents reporting bias or problems with the analysis of questionnaires or food records cannot be decided on the basis of these data. However, the reporting bias explanation seems the more likely.

We have found that both variation of diet over time and intraindividual correlation of proximal questionnaires and food records have an influence on the estimates of these quantities. Although Beaton et al. (3) mentioned the potential importance of intraindividual correlation few authors have discussed its assessment or applied any corrections for it. Willett (5) reports not finding substantial interindividual correlations in his studies. However, of the 5 nutrients we have

investigated in the Women's Health Trial data, two displayed high intraindividual correlations, two moderate and one low. For the two nutrients with high correlations, the estimated regression slopes were reduced by between 0.14 and 0.23, i.e. by 13 to 25% of their previous values, when adjustment for intraindividual correlation was made.

Intraindividual correlation can only be estimated when repeat assessments using both questionnaire and food record are obtained. The frequent absence of repeat assessments may explain why investigators have not routinely considered such correlation. However we note that although the value of the correlation r cannot be estimated without repeated questionnaires, the adjustments of equations (4) and (5) can in fact be made with only one questionnaire together with two or more food records. This is because the quantity $r \cdot \sigma_{WF}\sigma_{WQ}$, the intraindividual covariance, may be estimated by the difference between covariance between the questionnaire with its proximal food record and the covariance between the questionnaire with the other non-proximal food records. The minimal data for applying the intraindividual correlation correction are one questionnaire and two food records, one of which is proximal to the questionnaire and the other distant in time. Should both food records be taken at a time distant from the questionnaire determination (say 6 months or more) then the likelihood of a substantial intraindividual correlation is less.

We recommend that nutritional epidemiologists using food questionnaire data employ in their validation exercises repeated assessments, and consider the problem of intraindividual correlation as well as the better known variation-in-diet problem. The statistical methods for dealing with these are described in the Appendix.

## ACKNOWLEDGEMENTS

We thank Ron Knickerbocker, David Pee and Lisa Licitra for assistance with the computer programming, and Carolyn Clifford and Gladys Block for valuable advice.

Carroll's research was supported by NIH Grant GM-39015 and partially completed during a visit to Purdue University's Center for Decision Sciences, which is supported by NSF Grants DMS-8702620 and DMS-8717799.

## REFERENCES

1. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. Am J Epidemiol 1985;122:51-65.

2. Beaton GH, Milner J, Corey P, et al. Sources of variance in 24-hour dietary recall data: implications for nutrition study, design and interpretation. Am J Clin Nutr 1979;32:2546-59.

3. Lui K, Stamler J, Dyer A, et al. Statistical methods to assess and minimize the role of intra-individual variability in observing the relationship between dietary lipids and serum cholesterol. J Chronic Dis 1978;31:399-418.

4. Salvini S, Hunter DJ, Sampson L, Stampfer MJ, Colditz GA, Rosner B, Willett WC. Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption. Int J Epidemiol 1989;18:858-67.

5. Willett WC. Nutritional epidemiology. New York: Oxford University Press, 1990, p 276.

6. Prentice RL, Pepe M, Self SC. Dietary fat and breast cancer: a quantitative assessment of the epidemiologic literature and discussion of methodologic issues. Cancer Res 1989;44:3147-56.

7. Freedman L, Schatzkin A, Wax Y. The effect of dietary measurement error on the sample size of a cohort study. Am J Epidemiol. In press.

8. Rosner, B., Willett, W.C. and Spiegelman, D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8:1051-70.

9. Henderson MM, Kushi LH, Thompson DJ, Gorbach SL, Clifford CK, Insull W, Moskowitz M, Thompson RS. 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 1990;19:115-33.

10. Gersovitz M, Madden JP, Smiciklas-Wright H. Validity of the 24 hour dietary recall and seven-day record for group comparisons. J Am Diet Assoc 1978;73:48-55.

11. Subcommittee on Criteria for Dietary Evaluation, National Research Council. Nutrient adequacy: assessment using food consumption surveys. Washington, DC: National Academy Press, 1986, p 52.

12. Block G, Hartman AM, Dresser CM, Carroll MD, Gannon J, Gardner L. A data-based approach to diet questionnaire design and teaching. Am J Epidemiol 1986;124:453-69.

13. Rosner B, Willett WC. Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. Am J Epidemiol 1988;127:337-86.

# APPENDIX

## Estimators and Standard Errors of Slope and "Bias" in a Measurement Error Model Relating Food Questionnaire to True Intake, Using Observations of Food Questionnaire and Food Records

We consider that for the $i^{th}$ individual ($i = 1, ..., n$) with true intake $T_i$, we have $m_1 + m_2$ replicate measures of the questionnaire $Q_{ij}$ ($j = 1, ..., m_1+m_2$) and $m_2 + m_3$ replicate measures of the food record $F_{ij}$ ($j = m_1+1, ..., m_1+m_2+m_3$). The $m_2$ questionnaires and food records with the same index j are measured close in time to each other. The analysis described below requires that $m_2$ and $m_3$ are both greater than zero. The model relating questionnaire and food record to the individual's true intake is as follows:

*Questionnaire Alone*

$$Q_{ij} = \alpha_{QT} + \beta_{QT}T_i + \upsilon_i + \varepsilon_{ij}; \qquad j = 1, ..., m_1.$$

where $\alpha_{QT}$ is the intercept, $\beta_{QT}$ is the slope, $\varepsilon_{ij}$ represents the variation due to replication of the questionnaire and $\upsilon_i$ the variation about the regression line for each individual. The terms $\varepsilon_{ij}$ and $\upsilon_i$ are mutually independent with mean 0 and variances $\sigma^2_\varepsilon$ and $\sigma^2_\upsilon$ respectively.

*Questionnaire and Food Record*

$$Q_{ij} = \alpha_{QT} + \beta_{QT}T_i + \upsilon_i + \varepsilon_{ij};$$

$$F_{ij} = T_i + u_{ij}; \qquad j = m_1+1, ... , m_1+m_2.$$

where $u_{ij}$ represent the variation due to replication of the food record. The terms $u_{ij}$ are mutually independent with mean 0 and variance $\sigma_u^2$. The terms $u_{ij}$ and $\varepsilon_{ij}$ have correlation $\rho$, but $u_{ij}$ and $\varepsilon_{ik}$ are independent when j is not equal to k.

*Food Record Alone*

$$F_{ij} = T_i + u_{ij}; \qquad j = m_1+m_2+1, \ldots, m_1+m_2+m_3.$$

In the example of the paper, $m_1 = 0$, $m_2 = 2$ and $m_3 = 1$.

To implement the moment estimation method of Fuller (1987, page 186), define the following for the $i^{th}$ individual:

$\overline{F}_i$ = sample mean of all food records;

$\hat{\sigma}^2_{u_i}$ = sample variance of all food records (using the denominator $m_2+m_3-1$)

$\overline{Q}_i$ = sample mean of all questionnaires;

$\hat{\sigma}_{\varepsilon u_i}$ = sample covariance between the $m_2$ questionnaires and food records measured at the same time (using the denominator $m_2-1$)[1].

---

[1] It may sometimes happen that $m_2$ equals 1, i.e. that only one questionnaire and food record are proximal. In that case $\hat{\sigma}_{\varepsilon u_i}$ may be calculated as

$$(Q_{i,m_1+1} - \overline{Q}_i)(F_{i,m_1+1} - \overline{F}_i) \Big/ \left[1 - \frac{1}{m_1+1} - \frac{1}{m_3+1} + \frac{1}{(m_1+1)(m_3+1)}\right].$$

Moreover if $m_2 = 1$ and $m_1 = 0$ then $\hat{\sigma}_{\varepsilon u_i}$ may still be calculated as

$$Q_{i,m_1+1}(F_{i,m_1+1} - \overline{F}_i) \Big/ \left[1 - \frac{1}{m_3+1}\right].$$

However the program assumes that $m_2 \geq 2$ and will require modification for these special cases.

Let

$$\hat{\Omega}_{u_i} = \begin{pmatrix} 0 & 0 \\ 0 & \hat{\sigma}^2_{u_i} \end{pmatrix}; \qquad \hat{\Omega}_{\varepsilon u_i} = \begin{pmatrix} 0 \\ \hat{\sigma}_{\varepsilon u_i} \end{pmatrix};$$

$$A_i = \begin{pmatrix} 1 & \overline{F}_i \\ \overline{F}_i & \overline{F}^2_i \end{pmatrix} - \frac{1}{m2+m3}\,\hat{\Omega}_{u_i}\ ; \text{and}$$

$$B_i = \begin{pmatrix} \overline{Q}_i \\ \overline{Q}_i\,\overline{F}_i \end{pmatrix} - \frac{m_2}{(m_1+m_2)(m_2+m_3)}\,\hat{\Omega}_{\varepsilon u_i}\,.$$

Then the parameter estimates are given by the solution to the estimating equation $\sum_{i=1}^{n}\left[B_i - A_i\begin{pmatrix}\alpha_{QT}\\ \beta_{QT}\end{pmatrix}\right] = 0$, and can be written

$$\begin{pmatrix}\hat{\alpha}_{QT}\\ \hat{\beta}_{QT}\end{pmatrix} = \left(\sum_{i=1}^{n} A_i\right)^{-1} \sum_{i=1}^{n} B_i.$$

If we define

$$C_i = B_i - A_i\begin{pmatrix}\hat{\alpha}_{QT}\\ \hat{\beta}_{QT}\end{pmatrix},$$

then $\sum_{i=1}^{n} C_i = 0$, and using Taylor's expansion of the estimating equation, the estimated covariance matrix of the parameter estimates is given by

$$\frac{n}{n-2}\left(\sum_{i=1}^{n} A_i\right)^{-1} \sum_{i=1}^{n} C_i C_i' \left(\sum_{i=1}^{n} A_i\right)^{-1},$$

where $C_i'$ denotes the transpose of matrix $C_i$.

Let $\mu_T$ be the population mean of the true nutrition intake $T_i$; $\mu_T$ is estimated by $\hat{\mu}_T = n^{-1}\sum_1^n \overline{F}_i$, the overall mean of the food record intakes. The "bias" is defined to be

$$b_{QT} = \alpha_{QT} + \mu_T(\beta_{QT} - 1),$$

this being the intercept for regressing $Q-\mu_T$ on $T-\mu_T$. The bias estimate is

$$\hat{b}_{QT} = \hat{\alpha}_{QT} + \hat{\mu}_T(\hat{b}_{QT} - 1).$$

We now show how to estimate the standard error of $\hat{b}_{QT}$.

Define

$$A_{i*} = \begin{pmatrix} A_i & 0 \\ 0 & 1 \end{pmatrix}; \qquad B_{i*} = \begin{pmatrix} B_i \\ \overline{F}_i \end{pmatrix};$$

$$C_{i*} = B_{i*} - A_{i*}\begin{pmatrix} \hat{\alpha}_{QT} \\ \hat{\beta}_{QT} \\ \hat{\mu}_T \end{pmatrix}; \qquad e_* = \begin{pmatrix} 1 \\ \hat{\mu}_T \\ \hat{\beta}_{QT}-1 \end{pmatrix};$$

and $$D_* = \left(\sum_{i=1}^n A_{i*}\right)^{-1} \sum_{i=1}^n C_{i*}C_{i*}' \left(\sum_{i=1}^n A_{i*}\right)^{-1}.$$

Since $\sum_1^n C_{i*} = 0$, $D_*$ is an estimate of the asymptotic covariance of $(\hat{\alpha}_{QT}, \hat{\beta}_{QT}, \hat{\mu}_T)$. An estimated standard error for $\hat{b}_{QT}$ is therefore:

$$\text{s.e.}(\hat{b}_{QT}) = \left(\frac{n}{n-2} e_*'D_*e_*\right)^{1/2}.$$

TABLE 1

*Values of the multiplication factor* $\left(1+\frac{\sigma^2_{WF}}{n_{F^\bullet}\sigma^2_{BF}}\right)$, *the estimated regression slopes* $\beta_{QF}$ *between two questionnaires and the average food record and the correlated slopes* $\beta_{QT}$, *assuming no intraindividual correlation, calculated from Women's Health Trial data*

| Nutrient | $1+\frac{\sigma^2_{WF}}{n_{F^\bullet}\sigma^2_{BF}}$ | 1 year questionnaire | | 2 year questionnaire | |
|---|---|---|---|---|---|
| | | $\beta_{QF}$ | $\beta_{QT}$ | $\beta_{QF}$ | $\beta_{QT}$ |
| Percent calories from fat | 1.32 | 0.82 | 1.08 | 0.64 | 0.84 |
| Log calories | 1.28 | 0.74 | 0.95 | 0.86 | 1.10 |
| Log protein | 1.32 | 0.69 | 0.91 | 0.61 | 0.81 |
| Log carbohydrate | 1.20 | 0.74 | 0.89 | 0.75 | 0.90 |
| Log saturated fat | 1.35 | 0.80 | 1.08 | 0.69 | 0.93 |

TABLE 2

*Estimates of the correlations $\rho_{QF}$ between two questionnaires and the average food record and the corrected correlation $\rho_{QT}$, assuming no intraindividual correlation, calculated from Women's Health Trial data*

| Nutrient | 1 year questionnaire | | 2 year questionnaire | |
|---|---|---|---|---|
| | $\rho_{QF}$ | $\rho_{QT}$ | $\rho_{QF}$ | $\rho_{QT}$ |
| Percent calories from fat | 0.60 | 0.69 | 0.50 | 0.57 |
| Log calories | 0.42 | 0.48 | 0.48 | 0.54 |
| Log protein | 0.41 | 0.47 | 0.36 | 0.41 |
| Log carbohydrate | 0.51 | 0.56 | 0.51 | 0.56 |
| Log saturated fat | 0.55 | 0.64 | 0.39 | 0.45 |

TABLE 3

*Estimates of intraindividual correlations (r) between questionnaire and food record estimated from Women's Health data*

| Nutrient | r |
|---|---|
| Percent calories from fat | 0.48 |
| Log calories | 0.25 |
| Log protein | 0.22 |
| Log carbohydrate | 0.03 |
| Log saturated fat | 0.61 |

TABLE 4

*Estimates of $\beta_{QT}$ and $\rho_{QT}$, taking account of intraindividual correlation, calculated from Women's Health Trial data*

| Nutrient | 1 year questionnaire | | 2 year questionnaire | |
|---|---|---|---|---|
| | $\beta_{QT}$ | $\rho_{QT}$ | $\beta_{QT}$ | $\rho_{QT}$ |
| Percent calories from fat | 0.94 | 0.60 | 0.70 | 0.48 |
| Log calories | 0.85 | 0.43 | 1.00 | 0.49 |
| Log protein | 0.81 | 0.43 | 0.71 | 0.37 |
| Log carbohydrate | 0.89 | 0.56 | 0.90 | 0.56 |
| Log saturated fat | 0.85 | 0.50 | 0.70 | 0.31 |

TABLE 5

*Estimates and standard errors, from the statistical model described in the Appendix, of slopes and "biases" of regression of questionnaire with true average intake, and correlation, using repeated questionnaire data*

| | Slope | S.E. | "Bias" | (S.E.) | Correlation (Questionnaire versus Truth) |
|---|---|---|---|---|---|
| Percent calories from fat | 0.83 | (0.13) | - 0.55 | (0.71) | 0.54 |
| Log calories | 0.93 | (0.21) | - 0.091 | (0.029) | 0.45 |
| Log protein | 0.76 | (0.22) | - 0.076 | (0.031) | 0.38 |
| Log carbohydrate | 0.89 | (0.13) | - 0.133 | (0.030) | 0.55 |
| Log saturated fat | 0.76 | (0.16) | - 0.104 | (0.040) | 0.46 |