

Binary Regression in Truncated Samples, with Application to Comparing Dietary Instruments in a Large Prospective Study

Douglas Midthune,^{1,*} Victor Kipnis,¹ Laurence S. Freedman,²
and Raymond J. Carroll³

¹Biometry Research Group, Division of Cancer Prevention, National Cancer Institute,
Executive Plaza North, Room 3131, 6130 Executive Boulevard,
MSC 7354, Bethesda, Maryland 20892-7354, U.S.A.

²Gertner Institute for Epidemiology and Health Policy Research,
Sheba Medical Center, Tel Hashomer 52161, Israel

³Department of Statistics, Texas A&M University, TAMU 3143,
College Station, Texas 77843-3143, U.S.A.

**email*: dm76q@nih.gov

SUMMARY. We examine two issues of importance in nutritional epidemiology: the relationship between dietary fat intake and breast cancer, and the comparison of different dietary assessment instruments, in our case the food frequency questionnaire (FFQ) and the multiple-day food record (FR). The data we use come from women participants in the control group of the Dietary Modification component of the Women's Health Initiative (WHI) Clinical Trial. The difficulty with the analysis of this important data set is that it comes from a truncated sample, namely those women for whom fat intake as measured by the FFQ amounted to 32% or more of total calories. We describe methods that allow estimation of logistic regression parameters in such samples, and also allow comparison of different dietary instruments. Because likelihood approaches that specify the full multivariate distribution can be difficult to implement, we develop approximate methods for both our main problems that are simple to compute and have high efficiency. Application of these approximate methods to the WHI study reveals statistically significant fat and breast cancer relationships when a FR is the instrument used, and demonstrate a marginally significant advantage of the FR over the FFQ in the local power to detect such relationships.

KEY WORDS: Biased sampling; Breast cancer; Case-control studies; Comparison of instruments; Measurement error; Misspecified models; Nutritional epidemiology; Truncation; Women's Health Initiative.

1. Introduction

This article is concerned with two questions of interest in nutritional epidemiology. The first is biological: is there a relationship between dietary fat intake and breast cancer? The second is methodological: how do various dietary assessment instruments compare in their power to detect diet-disease relationships? To address these questions we analyzed an important data set that was subject to biased sampling via truncation. This article addresses statistical issues involved in the analysis of such truncated data. Substantive results of the analysis may be found in Freedman et al. (2006).

Case-control studies, international comparisons, and laboratory experiments in animals generally support a positive association between fat consumption and the incidence of breast cancer (Howe et al., 1990). Conversely, a pooled analysis of several prospective studies, which are free of some of the biases that potentially affect case-control studies, has not found such an association (Hunter et al., 1996). A major problem with studies relating diet and disease is that of dietary measurement error. Both case-control and prospective studies em-

ploy self-reporting techniques for measuring dietary intake. For reasons of logistics and cost the most commonly used instrument is the food frequency questionnaire (FFQ) (Willet, 1990). Little is known about the nature and the extent of measurement error in FFQ-reported dietary fat intake, and there has been much discussion about whether such error could have led to the failure of the prospective studies to find a fat-breast cancer relationship (Prentice, 1996; Kipnis et al., 2001).

Other dietary assessment instruments are available, however, including multiple-day food records (FR) (see, e.g., Patterson et al., 1999). These instruments are more expensive to administer, but are often thought to be better measures of dietary intake than the FFQ, particularly because the respondent is required only to write down all that is eaten over a relatively short period, whereas completion of a FFQ requires estimation of average long-term diet, a much more difficult cognitive task. Moreover, the FR has already been used in a few large-scale studies (e.g., Bingham et al., 2003; Prentice et al., 2006). Nevertheless, the superiority of the FR over the

FFQ for studies of long-term diet and chronic disease is far from certain, because the FR measures only short-term intake, where long-term intake is of primary interest.

Recent calibration studies using reliable biomarker measures of protein and energy intake as reference instruments have indicated that measurement error in FFQ-reported protein and energy is substantial (Kipnis et al., 2003), and that for these two variables other dietary instruments, such as the FR or multiple 24-hour recalls, may have less measurement error than the FFQ and thus better predict true intake (Day et al., 2001; Schatzkin et al., 2003). No reliable reference instrument exists for dietary fat intake, however, so that direct estimation and comparison of measurement error in reported fat intake is not possible.

Recently, Bingham et al. (2003) reported the results of an indirect comparison of two instruments, a FFQ and a quantitative 7-day diary (a variant of the FR), both completed by a cohort of 13,070 women. They tested the fat–breast cancer hypothesis in this cohort and found that a statistically significant positive association between saturated fat intake and breast cancer incidence could be demonstrated using the 7-day diary, but not using the FFQ. These results suggested not only that the fat–breast cancer association exists, but that the 7-day diary may have more power than the FFQ to detect this association, thus providing indirect evidence that the 7-day diary may predict fat intake better than the FFQ.

One might object to this last conclusion, inasmuch as the measured association with breast cancer is distorted by measurement error and the positive association detected by the diary might be spurious; in order to properly compare the instruments, one must adjust the observed relationships for measurement error, but this is not possible without a reliable biomarker for fat intake. However, these objections may be overcome by noting that dietary measurement error predominantly causes simple attenuation of the estimated fat-intake regression coefficient, so that one may assume that tests of the null hypothesis, unadjusted for measurement error, are valid. Empirical evidence for this is available from a large biomarker study (Kipnis et al., 2003). In such a case, one may indeed compare the strength of observed relationships between fat and breast cancer for two instruments, and ascribe differences (if significant) to differences in the attenuation caused by the measurement error in the instruments. This reasoning underlies our current investigation.

The analysis of Bingham et al. (2003) was based on a relatively small number of breast cancer cases (168). Seeking to corroborate these results in a larger study, researchers at the National Cancer Institute and the Fred Hutchinson Cancer Research Center planned a similar comparison within the control group of the dietary modification (DM) arm of the Women's Health Initiative (WHI) Clinical Trial (Hays et al., 2003). The DM study is a randomized controlled trial of a low-fat diet that is high in fruits, vegetables, and grains. General eligibility criteria for the trial are provided in detail elsewhere (Hays et al., 2003). Women who participated in this trial completed both a FFQ and a 4-day FR on entry, allowing a comparison between these instruments similar to the comparison of Bingham et al.

The data were truncated in that approximately 42% of the women screened were excluded from the trial after the first

visit because they reported consuming a diet with less than 32% energy from fat, as estimated by the FFQ. Percent energy from fat is defined as $100 \times (\text{fat in kilo-calories}) / (\text{energy in kilo-calories})$. The purpose of this screening was to enroll into the study women having relatively high fat intake and thereby increase the difference in percent energy from fat between women in the dietary intervention and control groups.

The control group we analyzed comprised approximately 30,000 women who received general advice on diet and health, but no intensive dietary counseling. At the time of the analysis, after a median follow up of 7 years, 603 of these women had been diagnosed with invasive breast cancer. Staff at the WHI Clinical Coordinating Center selected two women with no diagnosis of breast cancer for every case, resulting in a sample of 603 cases and 1206 noncases. This was done because the cost of analyzing the food records of all women in the control group would have been prohibitive. Thus, not only is the sample truncated, but it is in the form of a nested case–control study.

The main goal of this article is to describe a simple methodology for estimating risks in the full (nontruncated) population using truncated data such as ours, both for the FFQ (on which the truncation is based) and for the FR. It is important for us to provide inference for the full population for two reasons. Firstly, the truncated population is of little biological interest because it is instrument specific. Secondly, the truncation, being based on one of the instruments, affects risk estimates differently for the two instruments and restricting inference to the truncated population would preclude a fair comparison of these two instruments.

The method uses as its motivation work by Gail, Wieand, and Piantadosi (1984), and requires little more than ordinary linear logistic regression. It posits a risk model in which the truncation variable is included, and then employs a new residualization method to approximate the marginal model when the truncation variable is not included. We will demonstrate that the approximation gives estimates with very small bias, and with good power.

The second goal of this article is to describe a simple methodology that allows comparison of the instruments in terms of the local power each would have in a nontruncated prospective study, again using data from the truncated sample. To do so, we develop estimates of the standard errors of the estimators that would have been obtained in a (hypothetical) nontruncated sample. We show that under some fairly mild assumptions, one can obtain an approximate comparison of the instruments without having to model the often high-dimensional distribution of all the covariates.

An outline of the article is as follows. Section 2 describes the potential for bias caused by truncated sampling and our method for addressing this problem. Section 3 describes the methods used to compare the two instruments when data are from a truncated sample. Section 4 presents results of simulations demonstrating that our approximate truncation-adjusted methods work well. Section 5 provides an analysis of the WHI study data. We find that various types of fat intake appear to be risk factors for breast cancer when using the FR to assess diet, but that no such effect is found using the FFQ. We also find that the estimated local power of the FR to detect a fat–breast cancer relationship is higher than that of

the FFQ and that the difference in local powers is marginally significant.

The first part of Section 6 points out some of the difficult asymptotic issues that arise in the analysis of nest case-control studies, as described in detail by Arratia, Goldstein, and Langholz (2005). The second part of Section 6 briefly describes two likelihood-based alternatives, comparing them with our methodology. Section 7 presents further discussion of our method.

2. Methods for Risk Analysis in Truncated Samples

Let Y be a binary response, let F be the risk factor of primary interest, and let X be all the other covariates. For example, F might be total fat intake as measured by a food record, and X might include energy (caloric) intake as measured by that food record, along with demographic and other risk factors such as age, race, education, body mass index, hormone usage, family history of breast cancer, biopsy for benign breast disease, alcohol consumption, and smoking status. In general, X can be high dimensional and complex.

The risk model of interest is a simple binary regression model

$$\text{pr}(Y = 1 | F, X) = H(\beta_0 + \beta_1 F + \beta_2^T X), \quad (1)$$

where $H(\cdot)$ is the logistic distribution function. Interest is in the logistic regression coefficient β_1 . Because of biased sampling, we observe only those subjects with $C > c_{\text{trun}}$, where C is a variable that may be related to Y , F , and X , and c_{trun} is the truncation cutoff value. We note that we are making an assumption that is unverifiable, at least within this truncated data set, namely that (1) holds in the full population. Models such as (1) are often assumed in cohort studies not subject to truncation.

2.1 The Role of the Truncation Variable C in Model (1)

The notation in model (1) is quite general, because the variables F and X are not specified, other than that F is of major interest, while X is usually high dimensional. There are three special cases:

- (i) The truncation variable C is itself the main variable of interest, in which case $F = C$ and we have the model

$$\text{pr}(Y = 1 | C, X) = H(\beta_0 + \beta_1 C + \beta_2^T X), \quad (2)$$

- (ii) C is not the main variable of interest but is included in the model anyway, in which case $X = (X_*, C)$ and the model becomes

$$\text{pr}(Y = 1 | F, X_*, C) = H(\beta_0 + \beta_1 F + \beta_{21}^T X_* + \beta_{22} C), \quad (3)$$

- (iii) C is part of neither F nor X .

As we will see in Section 2.2, models (2) and (3) are easily analyzed and the truncated sampling causes no bias.

2.2 Potential Bias Caused by Truncated Sampling

Truncated samples can always be used to estimate risks in the corresponding truncated population. Such an analysis can be valuable if the truncated population is of interest in its own right. In our example, however, the truncated population, women who report consuming $>32\%$ calories from fat on a

FFQ, is of little biological interest because it is instrument specific, depending on one of many error-prone dietary assessment methods. Moreover, because replications of the FFQ yield different results, the “truncated population” is not even defined within the instrument. For these reasons, we want to estimate risk in the full (untruncated) population.

If truncation variable C is a covariate in model (1), as in the special cases (2) and (3), then truncated sampling poses no issues in terms of consistency. If the model holds in the full population, then because the model is conditional on covariates that include C , it will hold in the truncated population as well. Of course, efficiency is affected by restricting the covariate space. Arratia et al. (2005, p. 911) show that even under complex sampling designs, such linear logistic models have an asymptotic theory that coincides with the usual population-based case-control studies.

If C is not a covariate in the model, however, then estimates obtained by regressing Y on (F, X) in the truncated sample will in general be inconsistent estimates of risk in the full population. To see this, let

$$S(F, X, \mathcal{A}) = \log \left\{ \frac{\text{pr}(C > c_{\text{trun}} | Y = 1, F, X)}{\text{pr}(C > c_{\text{trun}} | Y = 0, F, X)} \right\}, \quad (4)$$

where \mathcal{A} defines the conditional distribution of C given (F, X, Y) . It is easy to show that if model (1) holds in the full population, then within the truncated population (i.e., conditional on $C > c_{\text{trun}}$) the risk model is

$$\text{pr}(Y = 1 | X, F) = H\{\beta_0 + \beta_1 F + \beta_2^T X + S(F, X, \mathcal{A})\}. \quad (5)$$

In general, $S(F, X, \mathcal{A})$ will not be constant in (F, X) , and so model (1) will not hold in the truncated population, unless Y and C are conditionally independent given (F, X) . If one ignores the truncation and naively fits model (1) in the truncated sample, then in general the model will be misspecified and the resulting parameter estimates will be inconsistent estimates of the parameters in the full population.

Thus, one can safely ignore truncation only if Y and C are conditionally independent given (F, X) , which includes the special case when C is a covariate in the model. In the WHI study, one can ignore truncation when $F = \text{FFQ-reported percent energy from fat}$ (i.e., C), but not when $F = \text{FR-reported percent energy from fat}$, unless one assumes that breast cancer and FFQ-reported percent fat are conditionally independent given FR-reported percent fat. It is not thought that such conditional independence holds; moreover, assuming that it does hold is tantamount to assuming that the FR is a better instrument than the FFQ, and so is not appropriate for testing whether the FR is in fact better than the FFQ. Of course, even if the condition held, one could not perform a fair comparison of the instruments using the naive method, because truncation will affect the standard error of the FFQ estimate more than the standard error of the FR estimate. Thus special methods are needed for truncated samples.

2.2.1 Detailed example when C is not part of the model. As a specific example, suppose that C is normally distributed within the cases and controls as follows:

$$[C | Y = y, F, X] = \text{Normal}(\alpha_{0y} + \alpha_{1y} F + \alpha_{2y}^T X, \sigma_y^2), \quad (6)$$

for $y = 0, 1$; define $\mathcal{A} = (\alpha_{00}, \alpha_{10}, \alpha_{20}, \sigma_0, \alpha_{01}, \alpha_{11}, \alpha_{21}, \sigma_1)^T$.

Based upon the work of Satten and Kupper (1993), a sufficient condition for equation (6) to hold is that the following two conditions hold: (i) equation (6) holds among the controls $Y = 0$ and (ii) Y given (X, F, C) is linear logistic in (X, F, C) , with slope in C given as β_3 . Indeed, in that case, $\alpha_{10} = \alpha_{11} = \alpha_1$, $\alpha_{20} = \alpha_{21} = \alpha_2$, $\sigma_0^2 = \sigma_1^2 = \sigma^2$ but, in general, $\alpha_{00} \neq \alpha_{01}$, because $\alpha_{01} = \alpha_{00} + \sigma^2\beta_3$. Then

$$S(F, X, \mathcal{A}) = \log \left[\frac{1 - \Phi \left\{ (c_{\text{trun}} - \alpha_{00} - \beta_3\sigma^2 - \alpha_1F - \alpha_2^T X) / \sigma \right\}}{1 - \Phi \left\{ (c_{\text{trun}} - \alpha_{00} - \alpha_1F - \alpha_2^T X) / \sigma \right\}} \right]. \tag{7}$$

Here we see explicitly that if C is not part of the model of interest, and if the truncation is ignored, then fitting a linear logistic model in (F, X) is fitting a misspecified model, and hence there will be inconsistency. One practical point is that if $|\beta_3\sigma^2|$ is not large, then $S(F, X, \mathcal{A}) \approx 0$, and any large-sample bias will be small.

2.3 Estimation Methods for Truncated Samples

Our methodology consists of two approximations: (a) an approximate risk model based on a linear logistic model and (b) an approximate residualization.

2.3.1 Approximate risk model. Assume that the risk model for Y given (F, X, C) is linear logistic in (F, X, C) . As noted in Section 2.2, any regression model that includes the truncation variable C as a covariate will be immune from any large-sample bias arising from the truncated sample. Inclusion of C in the model, however, will change the regression parameters of the other covariates (F and X in our case), if C is correlated with them. To overcome this latter problem, we define $R = C - E(C|F, X)$, the residual of the regression of C on (F, X) . Supposing that we can estimate $E(C|F, X)$, we can then estimate R and include it as a covariate in the model. If $E(C|F, X)$ is linear in (F, X) , then Y is linear logistic in (F, X, R) , and R is uncorrelated with the other covariates. Furthermore, because by conditioning on (F, X, R) we are conditioning on C , the model is still immune from bias due to truncation.

Under our assumptions, Y has a linear logistic risk model in (F, X, R) :

$$\text{pr}(Y = 1 | F, X, R) = H(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}} + R\beta_{3,\text{trun}}), \tag{8}$$

where the subscript ‘‘trun’’ indicates that the risk model includes the truncation variable C . Note that because (8) includes C , it could be applied directly in the truncated sample if R were known or could be estimated. In the next section we describe how to estimate R .

Of course, we are not really interested in model (8), but instead want $\text{pr}(Y = 1 | F, X)$. More crucially, however, we want to be able to use the truncated data without having to directly account for truncation in our estimation. To do this, we fit model (8) to the data, use the approximation

$$\text{pr}(Y = 1 | F, X) \approx H(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}), \tag{9}$$

and then simply work with $(\beta_{1,\text{trun}}, \beta_{2,\text{trun}}) \approx (\beta_1, \beta_2)$.

Going from (8) to (9) is often justifiable because $R = C - E(C|F, X)$ is uncorrelated with (F, X) . The most direct case

justifying the approximation occurs when C given (F, X) is normally distributed in the population. If σ_R^2 is the variance of C given (F, X) , then by the usual probit approximation to the logistic,

$$\text{pr}(Y = 1 | F, X) \approx H \left\{ \frac{\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}}{(1 + \beta_{3,\text{trun}}^2\sigma_R^2/1.7^2)^{1/2}} \right\}. \tag{10}$$

When $\beta_{3,\text{trun}}\sigma_R$ is not large, the denominator inside (10) is close to 1.0, see Carroll et al. (2006, Chapter 4.8). Gail et al. (1984) give a more general calculation when R is independent of (F, X) , with a similar conclusion. Gail et al. also show that if regression function $H(\cdot)$ is an exponential rather than logistic function, then, because $\text{pr}(Y = 1 | F, X) = \exp(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}) \times E\{\exp(\beta_{3,\text{trun}}R) | F, X\}$, estimates based on (9) are consistent if R is independent of (F, X) , or, more generally, if $E\{\exp(\beta_{3,\text{trun}}R) | F, X\}$ is a constant. Under a rare disease assumption, the logistic function approximates an exponential function and logistic regression estimates based on (9) are approximately consistent. A Taylor series justification is given in the Appendix.

In summary, the method is to fit model (8) in the truncated sample, and interpret the coefficients for (F, X) as if they were the coefficients from the marginal model (1).

2.3.2 Approximate residual estimation. In order to implement this method, we must estimate $E(C|F, X)$ in the case-control study within the truncated sample. The obvious approach is to specify a model for C given (X, F) , and then use likelihood methods for truncated samples to estimate the regression function. In case-control studies, assuming a rare disease, this can be handled approximately by using the control data only. In our implementation, we assumed model (6) among the controls, fit the resulting linear model accounting for the truncation, and relied upon the rare disease assumption to obtain an approximately consistent estimate of $E(C|F, X)$ in the population. We then estimated residuals, $\hat{R} = C - \hat{E}(C|F, X)$, and substituted \hat{R} for R in (8). Standard errors were estimated using bootstrap methods.

3. Comparison of Instruments

In this section we consider how to compare two instruments in terms of their power to test hypotheses about β_1 when the data are from a truncated sample. Because truncation affects the power of the two instruments differently, a fair comparison has to evaluate how they would have performed had no truncation occurred.

3.1 General Technical Details

We begin with the simple binary regression model (1). Consider a sample of size n , and let $\mathcal{B} = (\beta_0, \beta_1, \beta_2^T)^T$ and let $\mathcal{Z} = (1, F, X^T)^T$. Using standard asymptotic theory, we have that

$$n^{1/2}(\hat{\mathcal{B}} - \mathcal{B}) \Rightarrow \text{Normal}\{0, \Sigma = (E[H(\mathcal{Z}^T\mathcal{B})\{1 - H(\mathcal{Z}^T\mathcal{B})\}]\mathcal{Z}\mathcal{Z}^T)^{-1}\}. \tag{11}$$

This means in particular that

$$n^{1/2}(\hat{\beta}_1 - \beta_1) \Rightarrow \text{Normal}\{0, \sigma_{\beta_1}^2 = (0, 1, 0^T)\Sigma(0, 1, 0^T)^T\}. \tag{12}$$

Using (12), local power is determined by the noncentrality parameter, namely

$$\Theta = n^{1/2}\theta = n^{1/2}\beta_1/\sigma_{\beta,1}. \quad (13)$$

Let $\Theta_{FR} = n^{1/2}\theta_{FR}$ and $\Theta_{FFQ} = n^{1/2}\theta_{FFQ}$ be the noncentrality parameters for the FR and FFQ, respectively, and let n_{FR} and $n_{FFQ} = \kappa n_{FR}$ be the respective sample sizes needed to attain a specified power to test $H_0: \beta_1 = 0$. Then κ , the relative sample size needed for the FFQ, is, asymptotically, $\kappa = (\theta_{FR}/\theta_{FFQ})^2$. Inference about the relative local power of the two instruments can be based on estimates of κ or $\Delta = \Theta_{FR} - \Theta_{FFQ}$, using bootstrap methods to estimate standard errors; we use Δ in our calculations, see Section 5 for discussion.

3.2 Estimating Noncentrality Parameters in Truncated Samples

The noncentrality parameter Θ in (13) is the asymptotic mean of $\hat{\beta}_1$ divided by the asymptotic standard deviation of $\hat{\beta}_1$ in unbiased samples of size n . As described in Section 2, biased sampling via truncation necessitates special methods for estimating β_1 , and the same is true for estimating $\sigma_{\beta,1}^2$. The purpose of the rest of this section is to propose simple methods and approximations that allow estimation of $\sigma_{\beta,1}^2$. We first suppose that we have a random sample from the truncated population, and reserve discussing the case-control issue until the end of Section 3.2.3.

3.2.1 Likelihood methods and their difficulties. To estimate $\sigma_{\beta,1}^2$ in general, we must estimate the expectation within equation (11). If one takes a likelihood approach to the problem, one has to specify a joint distribution for all the covariates (F, C, X) , in order to account for the truncated sampling. The complex nature of X , including continuous, ordered categorical and binary components, makes writing down this joint distribution unappealing at a practical level. It is far more appealing to construct an approximate method that estimates $\sigma_{\beta,1}^2$ without having to specify this high-dimensional distribution.

3.2.2 Simplification of equation (11). Let π_0 be the marginal probability of disease in the population. In most studies, the probabilities of disease $H(\mathcal{Z}^T\mathcal{B})$ vary much less than the covariates \mathcal{Z} themselves. This suggests replacing $H(\mathcal{Z}^T\mathcal{B})\{1 - H(\mathcal{Z}^T\mathcal{B})\}$ in (11) by $\pi_0(1 - \pi_0)$. Let σ_F^2 be the variance of F in the population, let its covariance with X be Σ_{FX} and let the covariance matrix of X be Σ_{XX} . With this approximation, (12) is easily seen to be

$$n^{1/2}(\hat{\beta}_1 - \beta_1) \approx$$

$$\text{Normal}\left[0, \sigma_{\beta,1,\text{approx}}^2 = \left\{\pi_0(1 - \pi_0)\left(\sigma_F^2 - \Sigma_{FX}\Sigma_{XX}^{-1}\Sigma_{FX}^T\right)\right\}^{-1}\right], \quad (14)$$

and consequently that

$$\Theta \approx n^{1/2}\beta_1/\sigma_{\beta,1,\text{approx}}. \quad (15)$$

Because π_0 in (14) does not depend on the instrument, one can estimate κ without having to estimate π_0 . For case-control studies, the approximation is the same, except that π_0 in (14) equals the proportion of cases in the study.

3.2.3 Estimation of $\sigma_{\beta,1,\text{approx}}^2$. Suppose that the bivariate pair (C, F) are jointly normally distributed given X , with

means linear in X and a constant covariance matrix. This implies the models

$$[F | C, X] = \alpha_0 + \alpha_1 C + \alpha_2^T X + \epsilon; \quad (16)$$

$$[C | X] = \gamma_0 + \gamma_1^T X + \eta, \quad (17)$$

where ϵ has mean zero and variance σ_ϵ^2 , while η has mean zero and variance σ_η^2 . It is easily shown that

$$\sigma_{\beta,1,\text{approx}}^2 = \left\{\pi_0(1 - \pi_0)\left(\sigma_\epsilon^2 + \alpha_1^2\sigma_\eta^2\right)\right\}^{-1}, \quad (18)$$

as demonstrated in the Appendix. If (F, X) is multivariate normal, then $\sigma_{\beta,1,\text{approx}}^2 \propto \{\text{var}(F | X)\}^{-1}$, although this is not generally true otherwise.

In order to implement (18), we need to estimate σ_ϵ^2 , α_1 , and σ_η^2 . This can be done either by joint maximum likelihood, or by the following simple device that we used in our implementation. First, estimate σ_ϵ^2 and α_1 by regressing F on (C, X) in the truncated sample. Then, estimate σ_η^2 by maximum likelihood using the likelihood based on model (17) and accounting for truncation,

$$f_{C,\text{trun}}(c | X, C > c_{\text{trun}}) = \frac{1}{\sigma_\eta} \phi\left\{\frac{c - \gamma_0 - \gamma_1^T X}{\sigma_\eta}\right\} \times \left[1 - \Phi\left\{\frac{c_{\text{trun}} - \gamma_0 - \gamma_1^T X}{\sigma_\eta}\right\}\right]^{-1}. \quad (19)$$

In the WHI controls study, estimation is complicated by the case-control sampling scheme. In Section 2.3.2, we described how one can estimate β_1 in case-control studies under a rare disease assumption. Similarly, one can estimate $\sigma_{\beta,1,\text{approx}}^2$ under this assumption by fitting (19) among the controls only.

4. Simulations

This section reports results of simulations designed to evaluate the performance of our truncation-adjusted procedure. To generate the data, we used an underlying measurement error model. However, the topic of this article is not measurement error modeling per se, and instead we focus on the effects of ignoring or accounting for the truncation on analyses that are done on the observed data.

In the simulations, nutrient data were generated to be similar to that for women in the Observing Protein and Energy Nutrition (OPEN) study (Kipnis et al., 2003), and for that reason are based on protein and energy, rather than fat and energy, intake. Table 1 presents the simulated distribution of true, FFQ- and FR-reported log intake of protein and energy, generated to be jointly normal, for three sets of simulations. In simulation 1, the FR is more closely correlated with true intake than is the FFQ; in simulation 2, the FFQ and FR are equally correlated with true intake; in simulation 3 (not shown in Table 1), the parameters for FFQ and FR in simulation 1 are reversed, so that the FFQ is more closely correlated with true intake than is the FR. The disease variable Y was generated from the logistic model

$$\begin{aligned} \text{pr}(Y = 1 | T_P, T_E) \\ = H[\alpha_0 + \alpha_1\{T_P - E(T_P)\} + \alpha_2\{T_E - E(T_E)\}], \end{aligned}$$

where T_P is true log intake of protein, T_E is true log intake of energy, $\alpha_0 = -4.2$, $\alpha_1 = 3.2$, and $\alpha_2 = 1.9$. These parameters

Table 1

Distribution of true and reported log intake of protein and energy in the simulation study, based on data from the OPEN study (Kipnis et al., 2003); in simulation 1, the 4-day FR is more closely correlated with true intake than is the FFQ; in simulation 2, the FFQ and FR are equally correlated with true intake.

Simulation	Parameter	True		FFQ		FR		
		Log protein	Log energy	Log protein	Log energy	Log protein	Log energy	
1	Mean	5.75	7.73	5.38	7.28	5.61	7.54	
	S.D.	0.19	0.16	0.42	0.39	0.24	0.21	
	Correlations							
	True log protein	1.00	0.49	0.25	0.05	0.49	0.15	
	True log energy		1.00	0.12	0.11	0.24	0.30	
	FFQ log protein			1.00	0.87	0.26	0.16	
	FFQ log energy				1.00	0.11	0.22	
	FR log protein					1.00	0.71	
	FR log energy						1.00	
2	Mean	5.75	7.73	5.38	7.28	5.38	7.28	
	S.D.	0.19	0.16	0.42	0.39	0.42	0.39	
	Correlations							
	True log protein	1.00	0.47	0.25	0.05	0.25	0.05	
	True log energy		1.00	0.12	0.11	0.12	0.11	
	FFQ log protein			1.00	0.87	0.23	0.15	
	FFQ log energy				1.00	0.15	0.20	
	FR log protein					1.00	0.87	
	FR log energy						1.00	

correspond to an overall disease risk of about $\lambda = 0.02$, a relative risk of about 4 comparing the 90th to the 10th percentile of true protein intake, and a relative risk of about 2 comparing the 90th to the 10th percentile of true energy intake.

Our data were generated as a nested-case control study. We generated a cohort of 30,000 subjects, and for each case randomly selected two controls. For each of our three simulations, we generated two types of study. In the first, there was no truncation. In the second, the cohort was generated with truncation, that is, all 30,000 members of the cohort had $C > c_{\text{trun}}$, where C is the logarithm of FFQ-reported percent energy from protein and $c_{\text{trun}} = \text{median}(C)$.

For each of the simulated studies, we fit logistic regression models,

$$\text{pr}(Y = 1 | Q_P, Q_E) = H(\beta_0 + \beta_1 Q_P + \beta_2 Q_E),$$

$$\text{pr}(Y = 1 | F_P, F_E) = H(\beta_0 + \beta_1 F_P + \beta_2 F_E),$$

where Q_P and Q_E are FFQ-reported log protein and log energy, and F_P and F_E are FR-reported log protein and log energy. Note that $C = Q_P - Q_E$. For the 1000 studies generated with nontruncated sampling, we fit the models using ordinary logistic regression. For the 1000 studies generated with truncated sampling, we fit the models using ordinary logistic regression (naive method) and the truncation-adjusted method described in Section 2.3. We also estimated $\sigma_{\beta_1}/n^{1/2}$ and Θ , what the standard error of β_1 and the noncentrality parameter would have been in a nontruncated sample, using the methods described in Section 3.2.

Table 2 presents the results of the simulations. The mean of $\hat{\beta}_1$, $\hat{\sigma}_{\beta_1}$, and $\hat{\Theta}$ in the nontruncated samples represent the true

parameters, that is, the expected values that one would have obtained in a nontruncated sample. For FFQ-reported intake in truncated samples, the naive and truncation-adjusted methods produced identical and unbiased estimates of β_1 ; this is because C is a linear function of the covariates Q_P and Q_E . However, the truncation-adjusted method estimators of σ_{β_1} and Θ were approximately unbiased for the FFQ, while the naive estimators were biased. In simulation 1, for example, the naive estimator of Θ was biased by 31%.

For FR-reported intake in truncated samples, the naive method produced biased estimates of β_1 , σ_{β_1} , and Θ . The bias of $\hat{\beta}_1$ was smallest (11%) in simulation 1 where the FR is a better instrument than the FFQ, and largest (24%) in simulation 3 where the FFQ is a better instrument than the FR. The truncation-adjusted method produced approximately unbiased estimates of the parameters for the FR in all three simulations.

We emphasize that the naive method for FFQ-reported intake in truncated samples produced unbiased estimates of β_1 in our simulations only because C was a function of the covariates in the model. In general the naive method will produce biased estimates, for example, when Q_P but not Q_E is a covariate in the model, or when C is the logarithm of FFQ-reported percent energy from fat and the covariates include saturated fat and energy but not total fat.

The results of simulation 2 show that the naive method cannot be used to compare the local power of dietary instruments in truncated samples. The mean of $\hat{\Theta}$, the estimated noncentrality parameter, was 3.61 for FFQ and 4.40 for FR when using the naive method, even though in simulation 2

Table 2

Simulation results; logistic regression and noncentrality parameters in nontruncated and truncated samples, estimated using FFQ or 4-day FR; simulated means, standard errors, and standard deviations of estimated parameters; $\hat{\sigma}_{\beta,1}/n^{1/2}$ estimates the hypothetical standard deviation of $\hat{\beta}_1$ in a nontruncated sample of size n ; $\hat{\Theta} = n^{1/2}\hat{\beta}_1/\hat{\sigma}_{\beta,1}$.

Simulation	Dietary assessment instrument	Type of sample	Estimation method for truncated sample	Mean $\hat{\beta}_1$ (S.E.)	S.D. $\hat{\beta}_1$	Mean $\hat{\sigma}_{\beta,1}/n^{1/2}$	Mean $\hat{\Theta}$ (S.E.)
1	FFQ	Nontruncated		1.27 (0.01)	0.24	0.25	5.14 (0.03)
		Truncated	Naive	1.26 (0.01)	0.35	0.35	3.57 (0.03)
		Truncated	Adjusted	1.26 (0.01)	0.35	0.24	5.20 (0.05)
	FR	Nontruncated		1.88 (0.01)	0.29	0.29	6.38 (0.03)
		Truncated	Naive	1.67 (0.01)	0.30	0.30	5.60 (0.03)
		Truncated	Adjusted	1.87 (0.01)	0.33	0.28	6.58 (0.04)
2	FFQ	Nontruncated		1.28 (0.01)	0.24	0.25	5.16 (0.03)
		Truncated	Naive	1.28 (0.01)	0.35	0.35	3.61 (0.03)
		Truncated	Adjusted	1.28 (0.01)	0.35	0.24	5.25 (0.05)
	FR	Nontruncated		1.29 (0.01)	0.24	0.25	5.19 (0.03)
		Truncated	Naive	1.06 (0.01)	0.25	0.24	4.40 (0.03)
		Truncated	Adjusted	1.27 (0.01)	0.26	0.24	5.23 (0.03)
3	FFQ	Nontruncated		1.91 (0.01)	0.30	0.30	6.46 (0.03)
		Truncated	Naive	1.90 (0.01)	0.41	0.41	4.64 (0.03)
		Truncated	Adjusted	1.90 (0.01)	0.41	0.29	6.63 (0.05)
	FR	Nontruncated		1.29 (0.01)	0.26	0.25	5.19 (0.03)
		Truncated	Naive	0.98 (0.01)	0.25	0.25	3.96 (0.03)
		Truncated	Adjusted	1.27 (0.01)	0.27	0.24	5.25 (0.04)

the two instruments had the same relation to true intake and so the same local power.

5. Analysis of the WHI Data

We present results of an analysis of data from the control group of the DM arm of WHI trial, described in the introduction. The analysis is designed to test hypotheses about associations between dietary fat and breast cancer incidence, and to compare the power of the FFQ and FR to detect such associations. We fit model (1) using the naive and truncation-adjusted methods described in Section 2. The exposure of interest, F , is the logarithm of total fat, saturated fat, polyunsaturated fat, or monounsaturated fat, and its logistic regression coefficient is estimated from the FFQ or FR data. The vector of other risk factors, X , consists of the following variables: logarithm of energy intake, duration of follow-up, age at entry to study (in 5-year age groups), clinical center region (North–East, South, Mid–West, West), hormone use (never, ever), family history (missing, no, yes), and breast biopsy (missing, no, yes). The truncation variable C is the logarithm of FFQ-reported percent energy from fat.

Table 3 presents results of the analysis. We first discuss results for the truncation-adjusted method, which our simulations have shown to be superior, and later compare the results of this method to those of the naive method. When using the FR to assess diet, total fat, polyunsaturated fat, and monounsaturated fat are statistically significant risk factors for breast cancer incidence, with estimated logistic regression coefficients of 0.78 (S.E. = 0.27), 0.51 (S.E. = 0.18), and 0.71 (S.E. = 0.23), respectively, while saturated fat, with an

estimated logistic regression coefficient of 0.31 (S.E. = 0.19), is positively associated but not statistically significant. When using the FFQ to assess diet, none of the types of dietary fat are statistically significant risk factors, and the estimated logistic regression coefficients are smaller than those estimated using the FR.

When we compare the local power of the FR and FFQ using $\hat{\Delta}$, the estimated difference in noncentrality parameters in a hypothetical nontruncated sample, tests do not reach the formal 0.05 level of significance for any of the types of dietary fat; the tests for total fat ($\hat{\Delta} = 2.07$, S.E. = 1.18) and polyunsaturated fat ($\hat{\Delta} = 2.25$, S.E. = 1.26), however, are close to significant, with p-values of 0.08 and 0.07, respectively, suggesting that the FR has greater local power to detect these diet–disease relationships.

We also calculated $\hat{\kappa}$, the estimated ratio of sample sizes needed to attain the same statistical power, but because the denominator of this ratio is far from statistical significance the confidence intervals for the ratio are very wide, meaning that $\hat{\kappa}$ cannot be reliably estimated.

The naive and truncation-adjusted methods produced very similar estimates of β_1 when using the FR to assess diet. This is somewhat surprising, because in simulations the naive method produced downward-biased estimates. We note, however, that in the simulations the bias of the naive method decreased as the correlation of FR with true intake increased relative to that of the FFQ. The naive estimate was greater than the truncation-adjusted method estimate in 9.4% of the studies in simulation 1, while it was greater in only 2.6% of the studies in simulation 2, and 0.8% of those in simulation 3.

Table 3

Analysis of the control group of the DM arm of the WHI Clinical Trial; estimated logistic regression and noncentrality parameters for different types of dietary fat, estimated using FFQ or 4-day FR; $\hat{\sigma}_{\beta,1}/n^{1/2}$ estimates the hypothetical standard deviation of β_1 in a nontruncated sample of size n ; $\hat{\Theta} = n^{1/2}\hat{\beta}_1/\hat{\sigma}_{\beta,1}$, $\hat{\Delta} = \hat{\Theta}_{FR} - \hat{\Theta}_{FFQ}$. Standard errors estimated by bootstrap methods.

Nutrient	Dietary assessment instrument	Estimation method	$\hat{\beta}_1$ (S.E.)	$\hat{\sigma}_{\beta,1}/n^{1/2}$	$\hat{\Theta}$	$\hat{\Delta}$ (S.E.)
Total fat log(g/d)	FFQ	Naive	0.39 (0.42)			
		Adjusted	0.39 (0.42)	0.31	1.24 (1.33)	
	FR	Naive	0.79 (0.26)			
		Adjusted	0.78 (0.27)	0.24	3.31 (1.16)	2.07 (1.18)
Saturated fat log(g/d)	FFQ	Naive	-0.09 (0.29)			
		Adjusted	0.03 (0.29)	0.25	0.14 (1.17)	
	FR	Naive	0.29 (0.19)			
		Adjusted	0.31 (0.19)	0.18	1.68 (1.05)	1.54 (1.12)
Polyunsaturated fat, log(g/d)	FFQ	Naive	0.08 (0.22)			
		Adjusted	0.13 (0.23)	0.17	0.77 (1.29)	
	FR	Naive	0.50 (0.18)			
		Adjusted	0.51 (0.18)	0.17	3.02 (1.08)	2.25 (1.26)
Monounsaturated fat, log(g/d)	FFQ	Naive	0.59 (0.36)			
		Adjusted	0.52 (0.37)	0.29	1.82 (1.29)	
	FR	Naive	0.72 (0.22)			
		Adjusted	0.71 (0.23)	0.20	3.47 (1.13)	1.65 (1.16)

Thus, the fact that the naive estimate is similar to, or even greater than, the truncation-adjusted method estimate in the present analysis is consistent with other results that indicate that the FR is a better instrument than the FFQ.

6. Alternative Approaches and Comments on Nested Case-Control Studies

6.1 Comments on Nested Case-Control Studies

Nested case-control studies may have a difficult asymptotic theory (Arratia et al., 2005) because of the possibility of correlations induced within the case and control samples. For example, Arratia et al. (2005) discuss an antihypertensive drug study in which there might be changes in the distribution of drug types during the study, and state that “differences in treatment within the case and control populations would make the modeling of the covariates by a common distribution within each group untenable” (p. 872), thus possibly affecting asymptotic theory. They go on to show, however (p. 911), that under reasonable conditions on the stability of the covariate distributions, the standard asymptotic theory for case-control studies in linear logistic regression holds for nonintercept parameters. The methodology we have proposed in Section 2 is based upon ordinary linear logistic regression, and is thus robust to the types of difficulties described by Arratia et al.

In our example, the potential difficulties with possible non-i.i.d. sampling may well be not of much concern, because the covariates are measures of macronutrients in diet at baseline of the WHI controls study, with recruitment done over a relatively short period (4 years), in a large cohort of over 29,000. However, this issue is debatable, and may impact on the methods described in Section 6.2 below.

6.2 Alternative Approaches

There are at least two alternative likelihood-based approaches.

Suppose we start with the primary risk model (1) where C is not part of the risk model, and that we also assume model (6) for C . Then, using (4) and (5), it is easy to construct the prospective likelihood function for (Y, C) given $(X, F, C > c_{trun})$. When we evaluated this method in the simulations described in Section 4, the standard deviations of the estimated risk parameters (β_1, β_2) were 50% greater than those from our proposed truncation-adjusted method. The culprit appears to be the estimation of \mathcal{A} in the offset $S(X, F, \mathcal{A})$ of model (4): when this was fixed at the correct value of \mathcal{A} , very much better performance was observed.

A second alternative is to assume that in addition to (1) and (6), the risk model is also linear logistic in (X, F, C) . Then, as noted in Section 2.2.1, the slope parameters $(\alpha_{1y}, \alpha_{2y})$ and the variance σ_y^2 do not depend on y . In other words, in the approach of the last paragraph that did so poorly, we make the changes that $(\alpha_{1y}, \alpha_{2y})$ and the variance σ_y^2 do not depend on y . Once again, it is easy to construct the prospective likelihood function for (Y, C) given $(X, F, C > c_{trun})$. When we implemented this approach in our simulations, the results were almost equivalent to those of our method. Our method, however, has the advantage that issues of rejective sampling raised by Arratia et al. (2005) are more clear than they are for these likelihood-based methods.

7. Discussion

This article has described methodology for risk estimation and instrument comparison from data that arise from a truncated

sample. Naive analysis that ignores the truncation has the potential to lead to important biases in risk estimates. To account for the truncation, we use a simple methodology that expands the original risk model to include the residual of the regression of the truncation variable on the covariates of interest. Under certain conditions, estimated logistic regression parameters for the covariates of interest in this expanded model are approximately consistent estimates of the parameters in the original risk model. In simulations these estimates had very small bias.

An analysis of the control group of the DM arm of the WHI clinical trial using these methods found that total fat, polyunsaturated fat, and monounsaturated fat were statistically significant risk factors for breast cancer incidence when a FR was used to assess diet, but that no significant effect was found when a FFQ was used to assess diet. These results are generally consistent with those of an analysis of another study by Bingham et al. (2003) that found significant effects for total fat and saturated fat when using a 7-day diary to assess diet, but no significant effects when using a FFQ. In combination, these results provide important evidence of an association between dietary fat intake and breast cancer incidence.

Although the naive and truncation-adjusted methods yielded estimated regression coefficients for the FR data that were very similar, this was not the case for the FFQ data. Furthermore, when we extended our truncation-adjusted method to deal with categorical covariates, such as quintiles of total fat, it yielded odds ratio estimates that were substantially larger than the naive estimates, both for FFQ and for FR data.

We also described simple methods in truncated data situations for comparing the power of dietary assessment instruments, in this case the FFQ and the FR, to detect diet-disease associations. A brute-force approach to this would have required the estimation of the population-level distribution of all covariates, a daunting task given the truncated sampling. We showed that under certain conditions one can avoid this approach, and we developed methods that address the problem. In the WHI analysis, we found that the estimated local power of the FR to detect a fat-breast cancer relationship was higher than that of the FFQ, and that the difference was marginally significant for total fat and polyunsaturated fat. This suggests that it is important to consider further the question of what dietary assessment instrument should be used in large prospective cohort studies of diet and disease.

The OPEN study (Kipnis et al., 2003) indicated that a single FFQ measures percent energy from protein only moderately worse than four 24-hour recalls (Schatzkin et al., 2003). It has been speculated that: (i) similar results might hold for percent energy from fat and (ii) a 4-day FR might perform similarly to four 24-hour recalls. Such speculations might lead one to expect that a FFQ and 4-day FR would perform similarly for percent energy from fat. The results of the current analysis, and that of Bingham et al. (2003), suggest that at least one of these speculations is incorrect. In particular, we suspect that measurement error in FFQ-reported intake may well be greater for fat than it is for protein.

This study was conducted in the control group of the WHI DM trial of a low-fat eating pattern. Principal results of the

randomized comparison in this trial were reported very recently (Prentice et al., 2006) and showed a reduction in breast cancer risk in the intervention group that did not quite attain conventional statistical significance. The estimated relative risk (hazard ratio) between the intervention and comparison group was 0.91 with a 95% confidence interval 0.83 to 1.01. If we apply our log relative risk estimate of 0.78 derived for total fat from the food record (Table 3) to the average difference in total fat intake between intervention and control groups reported in the trial, then we project a reduction of 8.8% compared to the 9.1% reduction observed in the trial. It might be objected that our projection of 8.8% should be adjusted for the attenuation that occurs due to measurement error, and that if that were done, the projection would be greatly in excess of the 9.1% observed reduction. However, it is also known that the intervention group underreported their energy intake in greater degree than the control group, so the reported average difference in total fat intake was very likely overestimated. Unfortunately, because there is no reliable biomarker for fat intake, we can know neither the true average difference in total fat intake between the groups, nor the exact amount of attenuation in the relative estimate caused by measurement error. However, one can say that the results of our truncation-adjusted FR-based analysis are consistent with the trial results, within the bounds of current knowledge.

ACKNOWLEDGEMENTS

RJC's research was supported by a grant from the National Cancer Institute (CA-57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We thank the associate editor and two referees for many detailed and helpful comments that led to a major restructuring of the article.

REFERENCES

- Arratia, R., Goldstein, L., and Langholz, B. (2005). Local central limit theorems, the high order correlations of rejective sampling, and logistic likelihood asymptotics. *Annals of Statistics* **33**, 871–914.
- Bingham, S. A., Luben, R., Welch, A., Wareham, N., Khaw, K. T., and Day, N. (2003). Are imprecise methods obscuring a relation between fat and breast cancer? *Lancet* **362**, 212–214.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, 2nd edition. New York: Chapman and Hall CRC Press.
- Day, N., McKeown, N., Wong, M., Welch, A., and Bingham, S. (2001). Epidemiological assessment of diet: A comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium. *International Journal of Epidemiology* **30**, 309–317.
- Freedman, L. S., Potischman, N., Kipnis, V., Midthune, D., Schatzkin, A., Thompson, F. E., Troiano, R. P., Prentice, R., Patterson, R., Carroll, R. J., and Subar, A. F. (2006). A comparison of two dietary instruments for evaluating the fat-breast cancer relationship. *International Journal of Epidemiology* **35**, 1011–1021.

- Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71**, 431–444.
- Hays, J., Hunt, J. R., Hubbell, F. A., Anderson, G. L., Limacher, M., Allen, C., and Rossouw, J. (2003). The Women's Health Initiative: Recruitment, methods and results. *Annals of Epidemiology* **13**, S18–S77.
- Howe, G. R., Hirohata, T., Hislop, T. G., Iscovich, J. M., Yuan, J., Katsouyanni, K., Lubin, F. Marubini, E., Madan, B., Rohan, T., Toniolo, P., and Shunzhang, Y. (1990). Dietary factors and risk of breast cancer: Combined analysis of 12 case-control studies. *Journal of the National Cancer Institute* **82**, 561–569.
- Hunter, D. J., Spiegelman, D., Adami, H. O., Beeson, L., van den Brandt, P. A., Folsom, A. R., Fraser, G. E., Goldbohm, R. A., Graham, S., and Howe, G. R. (1996). Cohort studies of fat intake and the risk of breast cancer—a pooled analysis. *New England Journal of Medicine* **334**, 356–361.
- Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Schatzkin, A., Subar, A. F., and Carroll, R. J. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidemiology* **153**, 394–403.
- Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. P., Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003). The structure of dietary measurement error: Results of the OPEN biomarker study. *American Journal of Epidemiology* **158**, 14–21.
- Patterson, R. E., Kristal, A. R., Fels Tinker, L., Carter, R. A., Bolton, M. P., and Agurs-Collins, T. (1999). Measurement characteristics of the Women's Health Initiative Food Frequency Questionnaire. *Annals of Epidemiology* **9**, 178–187.
- Prentice, R. L. (1996). Measurement error and results from analytic epidemiology: Dietary fat and breast cancer. *Journal of the National Cancer Institute* **88**, 1738–1747.
- Prentice, R. L., Caan, B., Chlebowski, R. T., et al. (2006). Low-fat dietary pattern and risk of invasive breast cancer. The Women's Health Initiative randomized controlled Dietary Modification trial. *Journal of the American Medical Association* **295**, 629–642.
- Satten, G. A. and Kupper, L. L. (1993). Inferences about exposure-disease association using probability of exposure information. *Journal of the American Statistical Association* **88**, 200–208.
- Schatzkin, A., Kipnis, V., Subar, A. F., Midthune, D., Carroll, R. J., Bingham, S., Schoeller, D. A., Troiano, R. P., and Freedman, L. S. (2003). A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: Results from the biomarker-based OPEN study. *International Journal of Epidemiology* **32**, 1054–1062.
- Willett, W. C. (1990). *Nutritional Epidemiology*, Chapter 5. New York: Oxford University Press.

Received February 2006. Revised March 2007.
Accepted March 2007.

APPENDIX

A.1 Taylor Series Justification of Truncation-Adjusted Method

A simple Taylor series approximation also suggests that in many practical cases, the transition from (8) to (9) is reasonable. To see this, write $H^{(1)}(x) = H(x)\{1 - H(x)\}$ and $H^{(2)}(x) = H^{(1)}(x)\{1 - 2H(x)\}$. Let $f_R(r|F, X)$ be the density function of R given (F, X) . Then, by a Taylor series expansion, because $E(R|F, X) = 0$ by definition,

$$\begin{aligned} \text{pr}(Y = 1 | F, X) &= \int H(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}} + r\beta_{3,\text{trun}}) f_R(r|F, X) dr \\ &\approx H(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}) \\ &\quad + \beta_{3,\text{trun}} H^{(1)}(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}) \\ &\quad \times \int r f_R(r|F, X) dr \\ &\quad + (1/2)\beta_{3,\text{trun}}^2 H^{(2)}(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}) \\ &\quad \times \int r^2 f_R(r|F, X) dr \\ &= H(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}) \\ &\quad + (1/2)\beta_{3,\text{trun}}^2 H^{(2)}(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}) \\ &\quad \times \int r^2 f_R(r|F, X) dr \\ &\approx H(\beta_{0,\text{trun}} + F\beta_{1,\text{trun}} + X^T\beta_{2,\text{trun}}), \end{aligned}$$

the last step assuming that the effect of R given (F, X) is small.

A.2 Proof of (18)

Combine (16)–(17) into the model

$$F = \gamma_0 + \alpha_0\gamma_1 + (\gamma_1\alpha_1 + \gamma_2)^T X + \epsilon + \alpha_1\eta.$$

Set $\zeta = \gamma_1\alpha_1 + \gamma_2$. Then $\sigma_F^2 = \zeta^T \Sigma_{XX} \zeta + \sigma_\epsilon^2 + \alpha_1^2 \sigma_\eta^2$ and $\Sigma_{FX} = \zeta^T \Sigma_{XX}$, from which the result follows.