

# Seemingly Unrelated Measurement Error Models, with Application to Nutritional Epidemiology

Raymond J. Carroll,<sup>1,\*</sup> Douglas Midthune,<sup>2</sup> Laurence S. Freedman,<sup>3,4</sup> and Victor Kipnis<sup>2</sup>

<sup>1</sup>Department of Statistics, Texas A&M University, TAMU 3143, College Station, Texas 77843-3143, U.S.A.

<sup>2</sup>Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Executive Plaza North, Room 3124, 6130 Executive Boulevard, MSC 7354, Bethesda, Maryland 20892-7354, U.S.A.

<sup>3</sup>Department of Mathematics, Statistics, and Computer Science, Bar Ilan University, Ramat Gan, 52900 Israel

<sup>4</sup>Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, 52621 Israel

\*email: carroll@stat.tamu.edu

**SUMMARY.** Motivated by an important biomarker study in nutritional epidemiology, we consider the combination of the linear mixed measurement error model and the linear seemingly unrelated regression model, hence *Seemingly Unrelated Measurement Error Models*. In our context, we have data on protein intake and energy (caloric) intake from both a food frequency questionnaire (FFQ) and a biomarker, and wish to understand the measurement error properties of the FFQ for each nutrient. Our idea is to develop separate marginal mixed measurement error models for each nutrient, and then combine them into a larger multivariate measurement error model: the two measurement error models are seemingly unrelated because they concern different nutrients, but aspects of each model are highly correlated. As in any seemingly unrelated regression context, the hope is to achieve gains in statistical efficiency compared to fitting each model separately. We show that if we employ a “full” model (fully parameterized), the combination of the two measurement error models leads to no gain over considering each model separately. However, there is also a scientifically motivated “reduced” model that sets certain parameters in the “full” model equal to zero, and for which the combination of the two measurement error models leads to considerable gain over considering each model separately, e.g., 40% decrease in standard errors. We use the Akaike information criterion to distinguish between the two possibilities, and show that the resulting estimates achieve major gains in efficiency. We also describe theoretical and serious practical problems with the Bayes information criterion in this context.

**KEY WORDS:** Akaike information criterion; Bayes information criterion; Latent variables; Measurement error; Mixed models; Model averaging; Model selection; Nutritional epidemiology; Seemingly unrelated regression.

## 1. Introduction

The purpose of this article is to propose a multivariate linear measurement error model motivated by the idea of seemingly unrelated regressions (Zellner, 1962), and to show how this idea can be applied to an important biomarker study in nutritional epidemiology to improve precision of key parameter estimates. Our approach differs from the standard multivariate measurement error model (Fuller, 1987; Tosteson, Buonaccorsi, and Demidenko, 1998; Wang et al., 1998) in three key aspects: (a) the motivation; (b) the details of the model; and (c) the need to include model selection in the analysis in order to have any hope of achieving efficiency gains.

### 1.1 *The Problem of Relating Diet to Disease*

This article was motivated by the problem of relating diet to cancer, and in particular by a study at the National Cancer Institute, the Observing Protein and Energy Nutrition (OPEN) study. This is the first large experiment that measures nutrient intakes via the most common dietary measurement instrument, the food frequency questionnaire (FFQ), along with

another instrument, the 24-hour recall (24HR), and biomarkers for protein and energy intake.

To set the stage for this study, and our analysis, some background will be helpful. Much of the recent literature on the relationship between diet and cancer has been based on analytic epidemiologic studies using FFQs. A number of large prospective studies of this kind have failed to find a consistent relationship between dietary components (such as fat, fiber, and fruits and vegetables) and cancers of the breast, colon, or rectum (Hunter et al., 1996; Fuchs et al., 1999; Michels et al., 2000). This may be explained by a true lack of diet-cancer associations or, alternatively, by serious methodological limitations of the studies themselves, especially due to FFQ measurement error.

Over the years, investigators have recognized that the reported values from FFQs are subject to substantial error, both systematic and random, that can profoundly affect the design, analysis, and interpretation of nutritional epidemiologic studies (Beaton et al., 1979; Freudenheim and Marshall, 1988; Freedman, Schatzkin, and Wax, 1990). Dietary measurement

error often attenuates (biases toward one) the estimates of disease relative risks (RRs) and reduces statistical power to detect their significance. An important relation between diet and disease, therefore, may be obscured. These considerations have prompted many investigations into studying more carefully the properties of dietary assessment instruments, especially the FFQ, using unbiased reference biomarkers of dietary intake such as urinary nitrogen (UN) for protein and doubly labeled water (DLW) for energy. A review of work with reference biomarkers and many references are given by Kipnis et al. (2001, 2003).

1.2 Models and Parameters for Nutrient Intake

We begin with the issue of measuring nutrient intake of a dietary variable with existing reference biomarker. We let  $i$  denote individual and  $j$  denote repeat measurement. For specificity, let “ $N$ ” be the nutrient of interest (e.g., protein or energy), “ $T$ ” true usual intake, “ $Q$ ” the FFQ, “ $F$ ” 24HR, and “ $M$ ” the biomarker. In all our work, we transformed the data via logarithms. The measurement error model of Kipnis et al. (2001) for such a nutrient is

$$\begin{aligned} Q_{Nij} &= \beta_{N0}^Q + \beta_{N1}^Q T_{Ni} + r_{Ni}^Q + \epsilon_{Nij}^Q; \\ F_{Nij} &= \beta_{N0}^F + \beta_{N1}^F T_{Ni} + r_{Ni}^F + \epsilon_{Nij}^F; \\ M_{Nij} &= T_{Ni} + \nu_{Nij}, \end{aligned} \tag{1}$$

where  $Q_{Nij}$ ,  $F_{Nij}$ , and  $M_{Nij}$  are the nutrient intakes obtained on the  $i$ th person from the  $j$ th repeat of the FFQ, the 24HR, and the reference biomarker, respectively. The subscript  $i$  runs from 1 to  $n$ , the total number of persons in the study. The subscript  $j$  refers to repeats of the instrument. The random variables ( $r_{Ni}^Q$ ,  $r_{Ni}^F$ ) are called person-specific biases, because they determine how different people with the same diet report their diet differently. The variables  $\epsilon_{Nij}^Q$ ,  $\epsilon_{Nij}^F$ , and  $\nu_{Nij}$  represent random within-person variation. The sets of random variables ( $T_{Ni}$ ), ( $r_{Ni}^Q$ ,  $r_{Ni}^F$ ), ( $\epsilon_{Nij}^Q$ ), ( $\epsilon_{Nij}^F$ ), and ( $\nu_{Nij}$ ) are assumed to be mutually independent, although variables within the parentheses are allowed to be correlated. Interestingly, the first part of model (1) was proposed in a different context by Cochran (1968), building on the work of Pearson (1902); see Carroll (2003). If true nutrient  $T_N$  were observable, model (1) would be a linear mixed model. Of course, the whole point is that truth is unknown here.

Two important quantities are the focus of our work.

- For dietary measurement error that is nondifferential with respect to disease, the attenuation factor, denoted throughout by  $\lambda$ , is defined as the slope of the regression of true intake on the FFQ, i.e.,  $\lambda = \text{cov}(T_N, Q_N) / \text{var}(Q_N)$ . This quantity describes the biases that will occur in analysis of disease that treats the FFQ values as if they were true. Such analysis, instead of estimating the true log-relative risk, estimates the log-relative risk basically multiplied by  $\lambda$ . Just as important,  $\lambda$  plays a central role in post hoc power calculations, i.e., given the observed variability in the FFQ. The sample size in a post hoc calculation needed to detect a specific alternative hypothesis is proportional to  $\lambda^{-2}$ .
- The second important quantity is the correlation of true intake and the FFQ, denoted throughout by  $\rho$ . Before designing a study, the sample size necessary to detect an

alternative hypothesis with given power is proportional to  $\rho^{-2}$ . Hence, designing a study sample size almost effectively reduces to estimating  $\rho$ .

1.3 Seemingly Unrelated Measurement Error Models

For any study of a given sample size, the difficulty is to obtain precise estimates of  $\lambda$  and  $\rho$ . While more data, in the form of more people, are always desirable, our goal is to obtain more precise estimates of these quantities via recognition that there are other data available, namely data on other nutrients.

Consider, for example, protein intake as the nutrient of interest. As is well known, protein intake is rather closely related to intake of total energy, and this relation suggests the possibility of using energy data to improve the precision of estimates for protein intake. This idea of *seemingly unrelated regressions* takes its measurement error form by simultaneously fitting models (1) for protein and energy intake. We will show in simulations and in our example (Sections 4 and 5) that, using this approach, we do see considerable gains in efficiency compared to using model (1) for protein only, with decreases in estimated standard errors by 40% and more. The problem is that this simple simultaneous measurement error model for protein and energy may not adequately describe the relation between the reported and true intakes of those nutrients. In general, it is possible that the FFQ and 24HR reports on each nutrient of interest (protein or energy) may be influenced by both true protein and true energy intakes. Extending model (1) to reflect this possibility and combining models for protein and energy lead to the following full simultaneous measurement error model

$$\begin{aligned} Q_{Pij} &= \beta_{P0}^Q + \beta_{P1}^Q T_{Pi} + \beta_{P2}^Q T_{Ei} + r_{Pi}^Q + \epsilon_{Pij}^Q; \\ F_{Pij} &= \beta_{P0}^F + \beta_{P1}^F T_{Pi} + \beta_{P2}^F T_{Ei} + r_{Pi}^F + \epsilon_{Pij}^F; \\ M_{Pij} &= T_{Pi} + \nu_{Pij}. \\ Q_{Eij} &= \beta_{E0}^Q + \beta_{E1}^Q T_{Pi} + \beta_{E2}^Q T_{Ei} + r_{Ei}^Q + \epsilon_{Eij}^Q; \\ F_{Eij} &= \beta_{E0}^F + \beta_{E1}^F T_{Pi} + \beta_{E2}^F T_{Ei} + r_{Ei}^F + \epsilon_{Eij}^F; \\ M_{Eij} &= T_{Ei} + \nu_{Eij}. \end{aligned} \tag{2}$$

In equations (2) and (3), ( $T_{Pi}$ ,  $T_{Ei}$ ), ( $r_{Pi}^Q$ ,  $r_{Pi}^F$ ,  $r_{Ei}^Q$ ,  $r_{Ei}^F$ ), ( $\epsilon_{Pij}^Q$ ,  $\epsilon_{Eij}^Q$ ), ( $\epsilon_{Pij}^F$ ,  $\epsilon_{Eij}^F$ ), ( $\nu_{Pij}$ ), and ( $\nu_{Eij}$ ) are assumed mutually independent. One can fit model (2) and (3), but a difficulty emerges, namely that in both our data and simulations the precision of estimates for the protein parameters is improved essentially not at all by the use of energy data. Thus, looked at in this way, we have increased the complexity of the modeling without gaining anything. This can be shown theoretically to be the case in the measurement error model in certain contexts (see Appendix A.1) and is similar to the results for the seemingly unrelated regressions on the same covariates as we discuss below. Thus, the full simultaneous measurement error model does not improve the precision of estimated parameters, and only the reduced model that sets  $\beta_{P2}^Q = \beta_{E1}^Q = \beta_{P2}^F = \beta_{E1}^F = 0$  by assuming that reported intakes of a nutrient are affected only by the true intake of that nutrient and not by addition of other nutrients, produces a gain in efficiency.

The problem, of course, is that the reduced model is not guaranteed to hold. Our simulations show that if we fit the reduced model when it does not hold, then negative biases in

attenuation  $\lambda$  and correlation  $\rho$  are as large as 25% in realistic contexts. This is unfortunate, because negative biases mean underestimation, and this level of underestimation suggests that the instruments are far worse than they actually are.

The obvious remedy to this issue is to use model selection methods. We base our methods on the Akaike information criterion (AIC), using a form of model averaging similar to that in the Bayesian literature. In our simulations, this method appears to resolve the issue of model robustness while still achieving efficiency gains when the reduced model holds.

#### 1.4 Outline

An outline of the article is as follows. Section 2 sets out the general framework of seemingly unrelated linear measurement error models. In Section 3 we describe the methods we develop. Section 4 gives results of a simulation study, which suggest that the weighted AIC method is nearly adaptive: it has seemingly unrelated measurement error models efficiency gains when the reduced model holds without the large biases that a reduced model analysis would incur when the full model holds. Section 5 describes the results of the OPEN study, which are basically in line with the simulation study. Section 6 has concluding remarks.

Theoretical derivations are given in an Appendix. They show that (a) there are no seemingly unrelated measurement error models efficiency gains in the full model (Section A.1); and (b) fitting the reduced model when the full model holds leads to asymptotic biases (Section A.2). In Section A.3, we derive distribution theory for AIC using work of Hjort and Claeskens (2003). In Section A.4, we show why one would expect to see problems with the Bayes information criterion (BIC), even though it is nominally a consistent model selector.

## 2. General Modeling Considerations

The seemingly unrelated regressions model can be developed as follows. For the primary variable of interest, denoted by subscript  $P$ , one has a vector response  $Y_P$ , a predictor vector  $X_P$ , and other factors  $Z_P$  related through a mixed model

$$Y_P = \beta_{0P} + X_P \beta_P + Z_P b_P + \epsilon_P, \quad (4)$$

where  $b_P = \text{Normal}(0, \Sigma_{bP})$  and  $\epsilon_P = \text{Normal}(0, \Sigma_{\epsilon P})$  are independent. Note that the subscript  $P$  can also stand for “protein” in our example. However, there is an extra variable, denoted by subscript  $E$  (“energy” in our example), for which a similar model holds

$$Y_E = \beta_{0E} + X_E \beta_E + Z_E b_E + \epsilon_E, \quad (5)$$

where  $b_E = \text{Normal}(0, \Sigma_{bE})$  and  $\epsilon_E = \text{Normal}(0, \Sigma_{\epsilon E})$ . To make the connection of this general model to our nutrient models (2) and (3), note the following. The terms  $Y_P$  and  $Y_E$  are the combination of FFQ and 24HRs for protein and energy, respectively. In both the full and reduced models  $Z_P b_P$  and  $Z_E b_E$  refer to the person-specific biases (the  $r$ 's) for protein and energy, respectively. However, the components  $X_P$  and  $X_E$  of (4) and (5) may differ depending on whether one is in the reduced model or in the full model. In the former case,  $X_P$  and  $X_E$  refer to true protein and energy intakes, respectively. In the latter case,  $X_P = X_E$  refers to all true protein and energy intakes taken together.

In the linear regression case that  $(X_P, X_E)$  are observed and random effects  $b_P = b_E = 0$ , Zellner (1962) showed that both (4) and (5) simultaneously may result in better estimates of the fixed effects  $\beta_P$  than simply fitting (4) alone. One case that this occurs is  $\Sigma_{\epsilon P} = \sigma_{\epsilon, P}^2 \times I$ ,  $\Sigma_{\epsilon E} = \sigma_{\epsilon, E}^2 \times I$ ,  $X_P \neq X_E$ , and the corresponding elements of  $\epsilon_P$  and  $\epsilon_E$  have common nonzero correlation. The condition that  $X_P \neq X_E$  makes it appear as if the regressions (4) and (5) were unrelated, and hence the term “*seemingly unrelated regressions*.” In our context,  $X_P = X_E$  for the full model, but because neither are observed, we found it surprising that for the full model, the seemingly unrelated regression phenomenon does not occur.

Our starting point of mixed models for (4) and (5) is vital in the measurement error context, where inference is not conditional on the random variables  $(X_P, X_E)$  and variances and covariances have a major role to play, since the crucial parameters  $\lambda$  and  $\rho$  are based on them. Indeed, in many problems, the variables  $X_P$  and  $X_E$  are not observable, and instead we observe

$$W_P = X_P + U_P; \quad (6)$$

$$W_E = X_E + U_E, \quad (7)$$

where the measurement errors  $(U_P, U_E)$  have mean zero, marginal covariance matrices  $\Sigma_{uuP}$  and  $\Sigma_{uuE}$ , and cross covariance matrix  $\Sigma_{uuPE}$ . In our context, for the reduced model  $W_P$  and  $W_E$  refer to biomarker protein and energy intakes, respectively. In the full model,  $W_P = W_E$  refers to the collection of all biomarker protein and energy intakes taken together.

Models (4) and (6) for the primary (protein) variable thus form a marginal linear mixed measurement error model as described in Wang et al. (1998) and Tosteson et al. (1998). Our idea is to combine this measurement error model with the seemingly unrelated measurement error model in the extra (energy) variable as given by (5) and (7).

## 3. Methods of Estimation Following Model Selection

The purpose of this section is to define and discuss the methods of estimation that we use in our study. Here, we will use the generic term  $\theta$  for the parameters in the model, while derived variables such as the attenuation  $\lambda$  and  $\rho$  are functions  $\mu(\theta)$ .

### 3.1 Full Model Estimate

The full model estimate is denoted by  $\hat{\theta}_F$ . This estimate is consistent and asymptotically normally distributed whether the full or reduced model holds. However, for the most part this estimate is little different from the univariate estimates, and thus does not take advantage of the seemingly unrelated measurement error models (SUMEM) phenomenon. Indeed, in Section A.1 we show that if all instruments have  $k$  replicates ( $j = 1, \dots, k$ ), only the FFQ and the biomarker are considered and there are no missing data, then the full model estimates of all model parameters exactly equal the univariate estimates. Clearly, this means that simply writing down a bigger model does not necessarily mean that the seemingly unrelated regression phenomenon will come into play.

### 3.2 Reduced Model Estimate

The reduced model estimate will be denoted by  $\hat{\theta}_R$ . One might expect, and we observe in our simulations, that if the full

model holds, then the reduced model estimate will generally be inconsistent, i.e., asymptotically biased. In Section A.2 we give a technical description of why this holds for the estimate of the attenuation  $\lambda$ .

3.3 Methods Based on AIC

AIC is defined as follows. Let  $L_F$  and  $L_R$  be the full model and reduced model log likelihoods, respectively, and let the number of parameters in those models be  $d_F$  and  $d_R$ . Then AIC is defined as  $AIC_F = -2L_F + 2d_F$  and  $AIC_R = -2L_R + 2d_R$ . The likelihood ratio test statistic is  $\mathcal{L} = 2(L_F - L_R)$ . The reduced model is chosen if  $AIC_R < AIC_F$ , i.e., if  $\mathcal{L}/2 - (d_F - d_R) < 0$ . The usual approach is to select the model on the basis of AIC, and then to compute the parameter estimate based on the selected model. Thus, if  $\hat{\theta}_F$  and  $\hat{\theta}_R$  are estimates of the parameter  $\theta$  in the full and reduced models, respectively, and if interest is in a derived parameter  $\mathcal{A} = \mu(\theta)$ , then a model selection estimate is

$$\hat{\mathcal{A}}_{ms} = I\{\mathcal{L}/2 - (d_F - d_R) > 0\}\mu(\hat{\theta}_F) + I\{\mathcal{L}/2 - (d_F - d_R) \leq 0\}\mu(\hat{\theta}_R). \tag{8}$$

The model selection estimate  $\hat{\mathcal{A}}_{ms}$  is not a smooth function of the log likelihood, and to rectify this one might use a weighted average of the full and reduced model estimates. Following Burnham and Anderson (1998), define the weights

$$\hat{\omega} = [1 + \exp\{\mathcal{L}/2 - (d_F - d_R)\}]^{-1}.$$

The weighted average AIC for the derived parameter  $\mathcal{A} = \mu(\theta)$  is given as

$$\hat{\mathcal{A}}_{aic} = \hat{\omega}\mu(\hat{\theta}_R) + (1 - \hat{\omega})\mu(\hat{\theta}_F). \tag{9}$$

It can be shown that if the full model holds then asymptotically AIC selects it with probability tending to 1, and full model inference is asymptotically correct. However, it turns out that AIC is not a consistent model selector, i.e., if the reduced model holds, then even in large samples AIC will select it with probability less than 1. Indeed, essentially by inspection one realizes that when the reduced model holds,

asymptotically it is selected with probability  $\text{pr}\{\chi^2(d_F - d_R) - 2(d_F - d_R) < 0\}$ , where  $\chi^2(d_F - d_R)$  denotes a chi-squared random variable with  $d_F - d_R$  degrees of freedom. For  $d_F - d_R = 1, 2, 3,$  and  $4$ , this probability is 0.85, 0.87, 0.88, and 0.91, respectively. The mean of  $\hat{\omega}$  in these cases is 0.63, 0.72, 0.78, and 0.82, respectively. Note that in our application,  $d_F - d_R = 4$ . Asymptotic theory based on the methods of Hjort and Claeskens (2003) is described in Section A.3.

3.4 Problems with Methods Based on BIC

We also investigated the use of BIC, for which  $d_F - d_R$  is replaced by  $(d_F - d_R)\log(n)/2$ . Unlike the AIC method, the BIC method is known to be a consistent model selection procedure, and this is often touted as a reason to prefer the method. However, in the simulations described in Section 4, we found that BIC was very badly biased for parameter estimation when the full model holds. The problem is not merely one that happens to occur in our simulation *setup*, but is, instead, a serious theoretical one. In the Appendix Section A.4, we describe a theory of full models as local alternatives to the reduced model where BIC will pick the reduced model with probability approaching one even when the full models holds, and hence the biases in the parameter estimates are to be expected. In situations such as ours, where the full and reduced models are not vastly different, BIC appears to be a poor choice as a model selection strategy.

4. Simulations

We performed a simulation study to understand the relative performance of the methods, namely, the univariate, full, reduced, selected AIC/BIC, and weighted AIC/BIC methods. We evaluated the performance of these methods in estimating the attenuation resulting from the FFQ measurement and its correlation with usual intake. Our simulations were designed to resemble the data collected in the OPEN study. Using the OPEN data for women after taking logarithms, described in detail in Section 5, we first fit models (2) and (3) to the protein and energy data using the reduced model. We then simulated from these models using the parameter estimates. The parameters that we used are given in Table 1.

Table 1

Reduced model fit for women, OPEN study, using the logarithm of protein intake and the logarithm of energy intake. Here “Estimate” refers to parameter estimate, while “SE” refers to its estimated standard error. In all simulations, for the reduced model,  $\beta_{P2}^Q = \beta_{E1}^Q = \beta_{P2}^F = \beta_{E1}^F = 0.0$ , while for the full model  $\beta_{P2}^Q = \beta_{E1}^Q = \beta_{P2}^F = \beta_{E1}^F = 0.2$ . By definition, the biomarker measurement errors are uncorrelated, so that  $\text{cov}(\nu_P, \nu_E) = 0$ . In addition,  $0 = \text{cov}(\epsilon_P^Q, \epsilon_E^F) = \text{cov}(\epsilon_P^Q, \epsilon_P^F) = \text{cov}(\epsilon_E^Q, \epsilon_P^F) = \text{cov}(\epsilon_E^Q, \epsilon_E^F)$ . The parameters were derived by fitting the reduced model to the OPEN data for women.

Estimate	SE	Estimate	SE	Estimate	SE			
$\beta_{P0}^Q$	2.21	0.69	$\beta_{E0}^Q$	5.19	0.80	$\beta_{P0}^F$	1.98	0.78
$\beta_{E0}^F$	4.36	0.81	$\beta_{P1}^Q$	0.55	0.12	$\beta_{E2}^Q$	0.27	0.10
$\beta_{P1}^F$	0.63	0.14	$\beta_{E2}^F$	0.41	0.10	$\text{var}(r_P^Q)$	0.114	0.014
$\text{var}(r_E^Q)$	0.112	0.013	$\text{var}(r_P^F)$	0.031	0.011	$\text{var}(r_E^F)$	0.032	0.008
$\text{var}(\epsilon_P^Q)$	0.048	0.005	$\text{var}(\epsilon_E^Q)$	0.039	0.004	$\text{var}(\epsilon_P^F)$	0.121	0.012
$\text{var}(\epsilon_E^F)$	0.079	0.007	$\text{cov}(r_P^Q, r_E^Q)$	0.102	0.013	$\text{cov}(r_P^Q, r_E^F)$	0.011	0.007
$\text{cov}(r_P^Q, r_P^F)$	0.014	0.008	$\text{cov}(r_E^Q, r_P^F)$	0.008	0.008	$\text{cov}(r_E^Q, r_E^F)$	0.017	0.007
$\text{cov}(r_P^F, r_E^F)$	0.025	0.008	$\text{cov}(\epsilon_P^Q, \epsilon_E^Q)$	0.037	0.004			
$\text{cov}(\epsilon_P^F, \epsilon_E^F)$	0.066	0.008	$\text{var}(T_P)$	0.036	0.007	$\text{var}(T_E)$	0.025	0.003
$\text{cov}(T_P, T_E)$	0.014	0.003	$\text{var}(\nu_P)$	0.042	0.005	$\text{var}(\nu_E)$	0.003	0.001

We considered four different cases. In all four cases, there were two FFQs, two 24HR measurements, and two protein biomarkers. In the first two cases, data were simulated from the reduced model with the parameters as described above. The true attenuations are 0.1165 and 0.3602 for protein and protein density, respectively, while the true correlations are 0.2531 and 0.4065. The sample sizes were  $n = 200$  and  $500$ . In these two cases, we assumed that the energy biomarker measurements were replicated only for a randomly selected subset of size  $m$  of the sample, these sample sizes being  $m = 25$  and  $m = 50$  for the two cases, respectively.

The other two cases were data generated from the full model, where now we set  $\beta_{P_2}^Q = \beta_{E_1}^Q = \beta_{P_2}^F = \beta_{E_1}^F = 0.20$ . This changed the attenuation and correlation. The true attenuations are 0.1329 and 0.2804 for protein and protein density, respectively, while the true correlations are 0.2889 and 0.2687.

#### 4.1 The Reduced Model

When the reduced model holds (cases 1 and 2 in Table 2), intuition plus the theory outlined in Section 3.3 argue that (a) the full model and the univariate protein model should be almost equivalent, (b) AIC should select the reduced model approximately 90% of the time, (c) the AIC weighted and selected methods should be essentially unbiased and also should be much more efficient than fitting the full model only, and (d) BIC will pick the reduced model essentially all the time. All these predictions are born out; see Table 2. BIC always, and AIC almost always picks the reduced model, and 30% decreases in standard errors are achieved by the AIC methods compared to fitting the full model or using protein alone.

#### 4.2 The Full Model

Intuition predicts here that the reduced model fit and the BIC methods (because they most often select the reduced model) will be badly biased, while the AIC methods, because

**Table 2**

*Results of protein intake simulations, with parameters set to be what we have estimated for the logarithm of protein intake. Cases 1 and 2 are for the reduced model, while cases 3 and 4 are for the full model. Here cases 1 and 3 have  $n = 200$  with substudy sample sizes  $m = 25$ , while cases 2 and 4 have  $n = 500$  and  $m = 50$ . The methods are as follows: Univariate is the univariate protein model (1), F-SUMEM is the full seemingly unrelated measurement error model for protein, R-SUMEM is the reduced seemingly unrelated measurement error model for protein, AIC selected is the model selection estimate (8), and AIC weighted is the weighted estimate (9). BIC selected and BIC weighted are similar to their AIC versions except BIC is used. For the selected methods, Weight is the percentage of the times that the reduced model holds. For the weighted methods, Weight is the mean weight for the reduced model over the simulations. Here % Bias is the percentage of bias in the parameter estimates, while RMSE is  $100 \times$  the square root of the mean squared error. If the % Bias column is blank, this means that the bias was  $< 2\%$ .*

Case	Model	Weight	Attenuation		Correlation	
			% Bias	RMSE	% Bias	RMSE
1 (Reduced, $n = 200$ )	Univariate			3.74		8.04
	F-SUMEM			3.74		8.04
	R-SUMEM			2.30		4.38
	AIC selected	0.90		2.68		5.41
	AIC weighted	0.81		2.59		5.14
	BIC selected	1.00		2.30		4.38
	BIC weighted	1.00		2.30		4.38
2 (Reduced, $n = 500$ )	Univariate			2.35		4.99
	F-SUMEM			2.35		4.99
	R-SUMEM			1.45		2.70
	AIC selected	0.90		1.66		3.26
	AIC weighted	0.82		1.63		3.19
	BIC selected	1.00		1.45		2.70
	BIC weighted	1.00		1.45		2.71
3 (Full, $n = 200$ )	Univariate			3.66		7.68
	F-SUMEM			3.66		7.68
	R-SUMEM			4.22	-26	8.81
	AIC selected	0.39	-7	3.98	-7	8.21
	AIC weighted	0.39	-8	3.72	-8	7.66
	BIC selected	0.95	-24	4.25	-23	8.86
	BIC weighted	0.93	-24	4.13	-22	8.58
4 (Full, $n = 500$ )	Univariate			2.35		5.01
	F-SUMEM			2.35		5.01
	R-SUMEM			3.97	-28	8.30
	AIC selected	0.04		2.42		5.13
	AIC weighted	0.07		2.42		5.12
	BIC selected	0.70	-17	3.57	-17	7.43
	BIC weighted	0.68	-17	3.35	-17	6.96

they reject use of the reduced model with high probability, will have smaller biases and lose little in the way of efficiency compared to the full model. All these predictions are born out; see cases 3 and 4 in Table 2.

### 5. Analysis of the OPEN Study

The OPEN study was conducted by the National Cancer Institute with 484 participants. A complete description of the study can be found elsewhere (Subar et al., 2003). Each participant was asked to complete an FFQ and 24HR on two occasions. Protein intake was also measured via the UN biomarker, with urine samples collected twice. All participants also had a single DLW biomarker measurement to estimate energy intake. In addition, the DLW measurement was repeated in 25 selected participants.

All the reported intake data and biomarker data on protein and energy were transformed to the logarithmic scale. Four methods of estimating the attenuation and correlation of the FFQ intake were investigated: univariate, full model, reduced model, and the AIC weighted method. The analyses were conducted for protein intake and energy intake. Calculations were also performed for protein density, which on the log scale is the difference between protein and energy. The attenuation and correlation estimates for protein density are therefore easily derived from the models (2) and (3).

For the univariate, full, and reduced models, standard error estimates are just the model-based estimates, that is, they are estimated from the inverse of the empirical information matrix. For the AIC weighted method, the standard errors are estimated using Burnham and Anderson's method (1998, p. 135, equation 4.11).

The AIC weights for the reduced model were 0.893 for males and 0.967 for females. Using the likelihood ratio test, the  $p$  values for the reduced model were 0.44 and 0.87 for males and females, respectively. Thus, there is little evidence that the reduced model is not reasonable.

Table 3 presents the results of the OPEN data, giving estimates and standard errors of the attenuation and correlation for the FFQ, according to the four methods. The results all show generally low (poor) attenuation (0.16 and below) and correlation coefficients (0.32 and below) for energy and protein, much lower than estimates from calibration studies that employ a self-reporting instrument as the reference instrument rather than the biomarkers used here (Kipnis et al., 2003).

As expected, the results for the univariate and full models are similar, as are the results for the reduced model and the AIC weighted method. Similar to those seen in the simulations, the standard errors from the reduced model and AIC weighted method are substantially reduced (by 30–44%) when compared to the full model. This has some useful benefits. First, planning of studies can be made with more certainty on the basis of the results from the AIC weighted method because of the extra precision. Second, comparisons of attenuations among different subgroups (such as males and females) can be made with more precision.

The results for protein density are different from that of protein and energy in two ways. First, the attenuations and correlations are substantially higher (better). Second, the reduction in standard error obtained from the reduced model and AIC weighted method is much smaller. That the atten-

uations and correlations are higher seems to emanate from the error structure of the FFQ. It seems that biases in energy and protein are large and highly correlated and that as a result biases in protein density are relatively small. That little reduction in standard error is achieved by the reduced model is explained by the much lower correlation between the variance components for protein density and energy, than between protein and energy. The implications of these findings are first that protein density seems to be more reliably determined by an FFQ than absolute protein or absolute energy intake, and that densities may therefore represent a better measure to use in trying to establish a link between dietary intake and disease.

### 6. Discussion

This article was motivated by the OPEN study in nutritional epidemiology. Our main approach is to combine a measurement error model for the variable of primary interest (protein) with a SUMEM for a variable of secondary importance (energy), in the hope of improving the precision of estimates of crucial importance, namely, the attenuation and the correlation with the truth. The result of our study is the class of SUMEM.

While the idea is simple the implementation is not. We have shown that simply building a bigger model may result in nothing more than extra complexity, without efficiency gain. Reducing the big model by setting some parameters to zero can, when the restrictions apply, result in SUMEMs achieving significant gains in efficiency, with 40% decrease in standard errors in our real and simulated data. On the other hand, we have shown both theoretically and in simulations that fitting the reduced model when the restrictions are false carries a price of a potentially large and practically important bias in parameter estimates.

In order to build model robustness, we have extended the SUMEM context by use of model selection and weighting. Model selection and model averaging were shown to provide the desired model robustness, and again significant gains in parameter estimation efficiency are achievable when the restrictions apply. However, the simulations and theory showed that BIC, while nominally a consistent model selection method, had serious biases in cases that the full model held.

The remaining problem that we have not addressed in this article is that of inference when using model selection. We have, however, a few remarks that readers may find of interest:

- Hjort and Claeskens (2003) provided a method for making such inference and constructing confidence intervals, one that we implemented in our simulations: we used the observed information matrix in our calculations when applying their method. The good news was that we found that their method gave nearly nominal coverage probabilities, but the bad news was that unfortunately the method was essentially the same as inference based upon the full model, and hence did not take advantage of the SUMEM effect; a similar conclusion in a much simpler, non-SUMEM context was discovered recently by Leeb and Pötscher (2005).
- We also implemented the method of Burnham and Anderson (1998) in our simulations. For the

**Table 3**

Results of the OPEN study. The methods are as follows: Univariate is the univariate protein model (1), Full is the full seemingly unrelated measurement error model for protein, Reduced is the reduced seemingly unrelated measurement error model for protein, and AIC weighted is the weighted estimate (9). For the univariate, full, and reduced models, standard error estimates are just the model-based estimates, that is, they are estimated from the inverse of the empirical information matrix. For the AIC-weighted method, the standard errors are estimated using Burnham and Anderson's method (1998, p. 135, equation 4.11). See Section 6 for a discussion that indicates their standard error estimates were reasonably accurate in our simulations.

Nutrient	Gender	Method	Attenuation		Correlation	
			Estimate	Standard Error	Estimate	Standard Error
Energy	Male	Univariate	0.080	0.025	0.199	0.061
		Full	0.079	0.025	0.195	0.061
		Reduced	0.058	0.016	0.143	0.036
		AIC weighted	0.061	0.017	0.149	0.040
	Female	Univariate	0.039	0.028	0.098	0.069
		Full	0.034	0.028	0.085	0.069
		Reduced	0.043	0.017	0.108	0.041
		AIC weighted	0.043	0.018	0.108	0.042
Protein	Male	Univariate	0.156	0.034	0.323	0.067
		Full	0.149	0.033	0.313	0.066
		Reduced	0.137	0.021	0.291	0.036
		AIC weighted	0.139	0.023	0.293	0.040
	Female	Univariate	0.137	0.041	0.298	0.088
		Full	0.129	0.041	0.282	0.088
		Reduced	0.114	0.024	0.250	0.048
		AIC weighted	0.114	0.025	0.251	0.049
Density	Male	Univariate	0.404	0.066	0.431	0.063
		Full	0.398	0.062	0.450	0.062
		Reduced	0.393	0.057	0.470	0.056
		AIC weighted	0.394	0.057	0.467	0.057
	Female	Univariate	0.316	0.084	0.346	0.087
		Full	0.321	0.080	0.373	0.088
		Reduced	0.328	0.074	0.389	0.075
		AIC weighted	0.328	0.074	0.389	0.076

AIC-weighted procedure, their method gave fairly accurate standard error estimates, which is one of the reasons why we have listed such standard error estimates in Table 3. The coverage probabilities attained by their method for the AIC-weighted procedure were often quite close to nominal, although we observed that in at least one case the actual coverage of a nominal 95% interval was 86%. Their method was less accurate for the AIC-selected procedure.

- Among the procedures we investigated, we do not recommend BIC in any form. Although it is nominally a consistent model selector, in practice both the selected and weighted forms of the procedure led to badly biased parameter estimates, and coverage probabilities using the Burnham and Anderson method were generally far below nominal. Leeb and Pötscher (2005) make similar remarks in their different context.

#### ACKNOWLEDGEMENTS

Carroll'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 Gerda Claeskens for showing us a preprint of her article.

#### REFERENCES

- Beaton, G. H., Milner, J., Corey, P., McGuire, V., Cousins, M., Stewart, E., de Ramos, M., Hewitt, D., Grambsch, P. V., Kassim, N., and Little, J. A. (1979). Sources of variance in 24-hour dietary recall data: Implications for nutrition study design and interpretation. *American Journal of Clinical Nutrition* **32**, 2546–2559.
- Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference*. New York: Springer.
- Carroll, R. J. (2003). Variances are not always nuisance parameters: The 2002 R. A. Fisher Lecture. *Biometrics* **59**, 211–220.
- Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics* **10**, 637–666.
- Freedman, L. S., Schatzkin, A., and Wax, Y. (1990). The impact of dietary measurement error on planning a sample size required in a cohort study. *American Journal of Epidemiology* **132**, 1185–1195.
- 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.
- Fuchs, C. S., Giovannucci, E. L., Colditz, G. A., Hunter, D. J., Stampfer, M. J., Rosner, B., Speizer, F. E., and Willett, W. C. (1999). Dietary fiber and the risk of colorectal

- cancer and adenoma in women. *New England Journal of Medicine* **340**, 169–176.
- Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.
- Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.
- Hunter, D. J., Spiegelman, D., Adami, H.-O., et al. (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., 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., 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.
- Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21–59.
- Michels, K. B., Giovannucci, E., Joshipura, K. J., Rosner, B. A., Stampfer, M. J., Fuchs, C. S., Colditz, G. A., Speizer, F. E., and Willett, W. C. (2000). Prospective study of fruit and vegetable consumption and incidence of colon and rectal cancers. *Journal of the National Cancer Institute* **92**, 1740–1752.
- Pearson, K. (1902). On the mathematical theory of errors of judgment. *Philosophical Transactions of the Royal Society of London A* **198**, 235–299.
- Subar, A. F., Kipnis, V., Troiano, R. P., et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The Observing Protein and Energy Nutrition (OPEN) study. *American Journal of Epidemiology* **158**, 1–13.
- Tosteson, T. D., Buonaccorsi, J. P., and Demidenko, E. (1998). Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. *Statistics in Medicine* **17**, 1959–1971.
- Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association* **93**, 249–261.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.
- mates of marginal parameters are the same as the marginal parameter estimates.
- The result is easily seen if we ignore the recall/records, so that only the FFQ and the biomarker measurements are available. Without loss of generality, we can assume that the true intakes have mean zero, and hence that  $\beta_{P0}^Q = \beta_{E0}^Q = 0$ . Then there are 15 unknown parameters, namely  $\text{var}(T_P)$ ,  $\text{var}(T_E)$ ,  $\text{cov}(T_P, T_E)$ ,  $\text{var}(\nu_{Pij})$ ,  $\text{var}(\nu_{Eij})$ ,  $\text{var}(r_{Pj}^Q)$ ,  $\text{var}(r_{Ej}^Q)$ ,  $\text{cov}(r_{Pj}^Q, r_{Ej}^Q)$ ,  $\text{var}(\epsilon_{Pi1}^Q)$ ,  $\text{var}(\epsilon_{Ei1}^Q)$ ,  $\text{cov}(\epsilon_{Pi1}^Q, \epsilon_{Ei1}^Q)$ ,  $\beta_{P1}^Q$ ,  $\beta_{P2}^Q$ ,  $\beta_{E1}^Q$ , and  $\beta_{E2}^Q$ . However, there are also 15 sufficient statistics, namely
- Terms that depend on the protein data alone:  $\text{var}(Q_{Pij})$ ,  $\text{cov}(Q_{Pi1}, Q_{Pi2})$ ,  $\text{var}(M_{Pij})$ ,  $\text{cov}(M_{Pi1}, M_{Pi2})$ , and  $\text{cov}(M_{Pi1}, Q_{Pi1})$ .
  - Terms that depend on the energy data alone:  $\text{var}(Q_{Eij})$ ,  $\text{cov}(Q_{Ei1}, Q_{Ei2})$ ,  $\text{var}(M_{Eij})$ ,  $\text{cov}(M_{Ei1}, M_{Ei2})$ , and  $\text{cov}(M_{Ei1}, Q_{Ei1})$ .
  - Terms that depend on combinations of protein and energy data:  $\text{cov}(Q_{Pi1}, Q_{Ei1})$ ,  $\text{cov}(Q_{Pi1}, Q_{Ei2})$ ,  $\text{cov}(M_{Pi1}, M_{Ei1})$ ,  $\text{cov}(M_{Pi1}, Q_{Ei1})$ , and  $\text{cov}(M_{Ei1}, Q_{Pi1})$ .
- In this sense, the full model is “just” identified. Consider the protein data. Since any marginal parameters, e.g., parameters associated only with the protein data such as the attenuation, are identifiable from sufficient statistics depending only on the protein data, the full model estimates and the marginal estimates must necessarily be the same.

## A.2 Bias of the Reduced Model When the Full Model Holds

The purpose of this section is to show that when one fits the reduced model even though the full model holds, the resulting measurement error model is misspecified. Since it is well known that incorrectly ignoring specified measurement error models leads to inconsistent estimates of attenuation and correlation in measurement error models, inconsistency of the reduced model estimates follows when the full model holds.

Recall the form of the full model for FFQs:

$$Q_{Pij} = \beta_{P0}^Q + \beta_{P1}^Q T_{Pi} + \beta_{P2}^Q T_{Ei} + r_{Pi}^Q + \epsilon_{Pij}^Q;$$

$$Q_{Eij} = \beta_{E0}^Q + \beta_{E1}^Q T_{Pi} + \beta_{E2}^Q T_{Ei} + r_{Ei}^Q + \epsilon_{Eij}^Q.$$

Consider first the case of marginal regression, i.e., using only the protein data to understand the error properties of the FFQ as it relates to protein. By adding and subtracting the regression of true protein on true energy, for some  $\beta_{P1*}^Q$ , we can rewrite the first equation as

$$\begin{aligned} Q_{Pij} &= \beta_{P0}^Q + \beta_{P1*}^Q T_{Pi} + \beta_{P2}^Q \{T_{Ei} - E(T_{Ei} | T_{Pi})\} + r_{Pi}^Q + \epsilon_{Pij}^Q \\ &= \beta_{P0}^Q + \beta_{P1*}^Q T_{Pi} + r_{Pi*}^Q + \epsilon_{Pij}^Q. \end{aligned} \quad (\text{A.1})$$

The major property of  $r_{Pi*}^Q$  defined in (A.1) is that it is uncorrelated with  $T_{Pi}$  and  $\epsilon_{Pij}^Q$ . Thus, marginally the FFQ for protein follows the model of Kipnis et al. (2001), and estimates of attenuation, correlation with usual intake, etc. are consistently estimated if one does not attempt to use the seemingly unrelated regression model.

A similar equation for energy intake also holds, for some  $\beta_{E1*}^Q$

$$Q_{Eij} = \beta_{E0}^Q + \beta_{E2*}^Q T_{Ei} + r_{Ei*}^Q + \epsilon_{Eij}^Q, \quad (\text{A.2})$$

where  $r_{Ei*}^Q = r_{Ei}^Q + \beta_{E1}^Q \{T_{Pi} - E(T_{Pi} | T_{Ei})\}$ .

Received December 2004. Revised April 2005.

Accepted April 2005.

## APPENDIX

### A.1 No Efficiency Gain in the Full Model

The purpose of this section is to show that if the full model holds, there are no missing data, and each instrument is replicated  $k$  times, so that  $j = 1, \dots, k$ , then the full model esti-

If, however, we combine (A.1) and (A.2) into a seeming unrelated “reduced” model, we have a major model violation. Specifically, it is not true that  $(T_{P_i}, T_{E_i})$  is independent of  $(r_{P_i^*}^Q, r_{E_i^*}^Q)$ , since, for example,  $T_{P_i}$  is not independent of  $r_{E_i^*}^Q$ . In other words, if we fit the reduced model when the full model holds, the reduced model is misspecified. Thus, it is no surprise that our simulations find that the reduced model fit leads to biased parameter estimates when the full model holds.

### A.3 Asymptotic Theory for the Weighted AIC Method

When the full model holds, AIC correctly chooses the full model with probability approaching 1 as  $n \rightarrow \infty$ , and hence both model selection and model averaging asymptotically result in an estimate equivalent to the full model estimate  $\hat{\lambda}_F$ . Since this estimate is asymptotically normally distributed, the bootstrap applies.

When the reduced model holds, a more complex argument is required. Indeed, the weighted AIC estimates are not even asymptotically normally distributed, nor does the bootstrap produce correct inferences (Hjort and Claeskens, 2003). To show the former fact, we follow Hjort and Claeskens (2003); our calculations are essentially equivalent to theirs, but because we are working in a simpler context our calculations are somewhat more transparent.

The general parameter is  $\theta$  that we partition as  $(\alpha^T, \gamma^T)^T$ . Under the reduced model, we have that  $\gamma = 0$ . We are interested in estimating a function of  $\alpha$  and  $\gamma$ , say  $\lambda = \mu(\alpha, \gamma)$ . In what follows,  $p$  will denote the number of components of  $\alpha$ , while  $q$  will denote the number of components of  $\gamma$ .

Let  $\hat{\theta}_F = (\hat{\alpha}_F^T, \hat{\gamma}_F^T)^T$  be the estimate computed using the full model, and let  $\hat{\theta}_R = (\hat{\alpha}_R^T, 0^T)^T$  be the estimate computed under the reduced model. We are interested in the parameter  $\lambda = \mu(\alpha, \gamma)$  under the reduced model when  $\gamma = 0$ .

The AIC weights are defined as follows. If the full model has  $d_F = p + q$  degrees of freedom and the reduced model has  $d_R = p$  degrees of freedom then the weight given to the reduced model is

$$\omega(\mathcal{L}_n) = \left[ 1 + \exp \left\{ \frac{\mathcal{L}_n}{2} - (d_F - d_R) \right\} \right]^{-1},$$

where  $\mathcal{L}_n$  is the likelihood ratio chi-squared statistic for testing the reduced model, i.e., whether  $\gamma = 0$ . Both the weighted AIC estimator and the model selection estimators can be written in the following form:

$$\begin{aligned} \hat{\lambda} &= \kappa(\mathcal{L}_n) \mu(\hat{\alpha}_R, 0) + \{1 - \kappa(\mathcal{L}_n)\} \mu(\hat{\alpha}_F, \hat{\gamma}_F); \\ \kappa(x) &= \omega(x) \quad (\text{weighted AIC}); \\ \kappa(x) &= I\{\omega(x) > 1/2\} \quad (\text{AIC model selection}). \end{aligned}$$

Let the information matrix for  $\theta$  be  $\mathcal{I}$  and its inverse  $\mathcal{I}^{-1}$  be partitioned as

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\alpha\alpha} & \mathcal{I}_{\alpha\gamma} \\ \mathcal{I}_{\gamma\alpha} & \mathcal{I}_{\gamma\gamma} \end{pmatrix}; \quad \mathcal{I}^{-1} = \begin{pmatrix} \mathcal{I}^{\alpha\alpha} & \mathcal{I}^{\alpha\gamma} \\ \mathcal{I}^{\gamma\alpha} & \mathcal{I}^{\gamma\gamma} \end{pmatrix}.$$

Note for example that  $\mathcal{I}^{\gamma\gamma} = (\mathcal{I}_{\gamma\gamma} - \mathcal{I}_{\gamma\alpha} \mathcal{I}_{\alpha\alpha}^{-1} \mathcal{I}_{\alpha\gamma})^{-1}$ ,  $\mathcal{I}^{\alpha\alpha} = (\mathcal{I}_{\alpha\alpha} - \mathcal{I}_{\alpha\gamma} \mathcal{I}_{\gamma\gamma}^{-1} \mathcal{I}_{\gamma\alpha})^{-1}$ , and that  $\mathcal{I}_{\alpha\alpha} = \{\mathcal{I}^{\alpha\alpha} - \mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} \mathcal{I}^{\gamma\alpha}\}^{-1}$ .

In what follows, all calculations are done under the reduced model. Let  $\mathcal{H}_\alpha$  be the likelihood score for  $\alpha$  in the reduced model, i.e., the derivative of the log likelihood with respect to  $\alpha$  and evaluated at  $\gamma = 0$ . Let  $\mathcal{H}_\gamma$  be defined similarly. By standard asymptotic approximations,  $n^{1/2}(\hat{\alpha}_R - \alpha) \approx \mathcal{I}_{\alpha\alpha}^{-1} n^{-1/2} \mathcal{H}_\alpha$  and

$$\begin{bmatrix} n^{1/2}(\hat{\alpha}_F - \alpha) \\ n^{1/2}\hat{\gamma}_F \end{bmatrix} \approx \mathcal{I}^{-1} \begin{pmatrix} n^{-1/2} \mathcal{H}_\alpha \\ n^{-1/2} \mathcal{H}_\gamma \end{pmatrix}.$$

This means that asymptotically,  $\hat{\alpha}_F$  is a linear function of  $\hat{\gamma}_F$  and  $\hat{\alpha}_R$ . To see this, replace approximation signs by equalities and note that  $n^{1/2}\hat{\gamma}_F = \mathcal{I}^{\gamma\alpha} n^{-1/2} \mathcal{H}_\alpha + \mathcal{I}^{\gamma\gamma} n^{-1/2} \mathcal{H}_\gamma$ . It thus follows that

$$\begin{aligned} n^{1/2}(\hat{\alpha}_F - \alpha) &= \mathcal{I}^{\alpha\alpha} n^{-1/2} \mathcal{H}_\alpha + \mathcal{I}^{\alpha\gamma} n^{-1/2} \mathcal{H}_\gamma \\ &= \mathcal{I}^{\alpha\alpha} n^{-1/2} \mathcal{H}_\alpha + \mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} \\ &\quad \times (\mathcal{I}^{\gamma\alpha} n^{-1/2} \mathcal{H}_\alpha + \mathcal{I}^{\gamma\gamma} n^{-1/2} \mathcal{H}_\gamma) \\ &\quad - \mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} \mathcal{I}^{\gamma\alpha} n^{-1/2} \mathcal{H}_\alpha \\ &= \{\mathcal{I}^{\alpha\alpha} - \mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} \mathcal{I}^{\gamma\alpha}\} n^{-1/2} \mathcal{H}_\alpha \\ &\quad + \mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} n^{1/2} \hat{\gamma}_F \\ &= n^{1/2}(\hat{\alpha}_R - \alpha) + \mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} n^{1/2} \hat{\gamma}_F, \quad (\text{A.3}) \end{aligned}$$

the last step following from the facts that  $\mathcal{I}_{\alpha\alpha} = \{\mathcal{I}^{\alpha\alpha} - \mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} \mathcal{I}^{\gamma\alpha}\}^{-1}$  and  $n^{1/2}(\hat{\alpha}_R - \alpha) = \mathcal{I}_{\alpha\alpha}^{-1} n^{-1/2} \mathcal{H}_\alpha$  asymptotically.

It is easily shown that  $\hat{\alpha}_R$  and  $\hat{\gamma}_F$  are asymptotically uncorrelated. Also, from standard calculations or by reference to the score test, the likelihood ratio test statistic is asymptotically equivalent to  $\mathcal{L}_n \approx n^{1/2} \hat{\gamma}_F^T (\mathcal{I}^{\gamma\gamma})^{-1} n^{1/2} \hat{\gamma}_F$ . Thus, the weight for the reduced model is asymptotically equivalent to

$$\begin{aligned} \omega_*(n^{1/2} \hat{\gamma}_F) \\ = \left[ 1 + \exp \left\{ \frac{n^{1/2} \hat{\gamma}_F^T (\mathcal{I}^{\gamma\gamma})^{-1} n^{1/2} \hat{\gamma}_F}{2} - (d_F - d_R) \right\} \right]^{-1}. \end{aligned}$$

We define  $\kappa_*(n^{1/2} \hat{\gamma}_F)$  in a similar fashion. Let  $\mu_\alpha$  and  $\mu_\gamma$  be the first partial derivatives of  $\mu(\alpha, \gamma)$  with respect to  $\alpha$  and  $\gamma$ , respectively. By direct calculation, remembering again that  $\lambda = \mu(\alpha, \gamma)$ , and using (A.3),

$$\begin{aligned} n^{1/2}(\hat{\lambda} - \lambda) &= \kappa_*(n^{1/2} \hat{\gamma}_F) n^{1/2} \{\mu(\hat{\alpha}_R, 0) - \mu(\alpha, 0)\} \\ &\quad + \{1 - \kappa_*(n^{1/2} \hat{\gamma}_F)\} n^{1/2} \{\mu(\hat{\alpha}_F, \hat{\gamma}_F) - \mu(\alpha, 0)\} \\ &= \mu_\alpha^T [\kappa_*(n^{1/2} \hat{\gamma}_F) n^{1/2} (\hat{\alpha}_R - \alpha) \\ &\quad + \{1 - \kappa_*(n^{1/2} \hat{\gamma}_F)\} n^{1/2} (\hat{\alpha}_F - \alpha)] \\ &\quad + \{1 - \kappa_*(n^{1/2} \hat{\gamma}_F)\} \mu_\gamma^T n^{1/2} \hat{\gamma}_F \\ &= \mu_\alpha^T [n^{1/2} (\hat{\alpha}_R - \alpha) + \{1 - \kappa_*(n^{1/2} \hat{\gamma}_F)\} \\ &\quad \times \mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} n^{1/2} \hat{\gamma}_F] \\ &\quad + \{1 - \kappa_*(n^{1/2} \hat{\gamma}_F)\} \mu_\gamma^T n^{1/2} \hat{\gamma}_F \\ &= \mu_\alpha^T n^{1/2} (\hat{\alpha}_R - \alpha) + \{1 - \kappa_*(n^{1/2} \hat{\gamma}_F)\} \\ &\quad \times (\mu_\gamma^T - \mu_\alpha^T \mathcal{I}_{\alpha\alpha}^{-1} \mathcal{I}_{\alpha\gamma}) n^{1/2} \hat{\gamma}_F, \quad (\text{A.4}) \end{aligned}$$

the last step following because  $\mathcal{I}^{\alpha\gamma} (\mathcal{I}^{\gamma\gamma})^{-1} = -\mathcal{I}_{\alpha\alpha}^{-1} \mathcal{I}_{\alpha\gamma}$ .

This means that if we let  $\mathcal{D}$  be normally distributed with mean zero and covariance matrix  $\mathcal{I}^{\gamma\gamma}$ , and if we let  $\mathcal{M}$  be normally distributed with mean zero and covariance matrix  $\mathcal{I}_{\alpha\alpha}^{-1}$ , and if  $\mathcal{D}$  and  $\mathcal{M}$  are independent, then the limit distribution of the weighted AIC class is the same as

$$n^{1/2}(\widehat{\lambda} - \lambda) \Rightarrow \mu_{\alpha}^T \mathcal{M} + \{1 - \kappa_*(\mathcal{D})\} (\mu_{\gamma}^T - \mu_{\alpha}^T \mathcal{I}_{\alpha\alpha}^{-1} \mathcal{I}_{\alpha\gamma}) \mathcal{D}. \quad (\text{A.5})$$

It is interesting to note that while  $n^{1/2}(\widehat{\lambda} - \lambda)$  has a limiting distribution, it is not a normal distribution.

#### A.4 Asymptotic Theory for the BIC Method

In contrast to AIC, the BIC weights are  $\omega(\mathcal{L}_n) = [1 + \exp\{\mathcal{L}_n/2 - \log(n)(d_F - d_R)/2\}]^{-1}$ . In our context, BIC is a consistent model selection procedure. To see this, note that under the reduced model,  $\mathcal{L}_n$  is asymptotically central chi-squared, so that  $\omega(\mathcal{L}_n) \rightarrow 1$  in probability. Under the full model,  $\mathcal{L}_n$  is asymptotically noncentral chi-squared with noncentrality parameter of order  $O(n^{1/2})$ , so that  $\omega(\mathcal{L}_n) \rightarrow 0$  in probability, as claimed. Since BIC is a consistent model se-

lection procedure, it follows trivially that the bootstrap is an asymptotically correct method of inference.

On the other hand, we have observed in our simulations (data not shown) that BIC leads to biased estimation when the full model holds. This empirical result, seemingly at odds with the theory described above, may be explained as follows. Consider local alternatives to the reduced model, i.e., a full model where  $\beta_{P_2}^Q$ ,  $\beta_{E_1}^Q$ ,  $\beta_{P_2}^F$ , and  $\beta_{E_1}^F$  all equal constants times  $n^{-1/2}$ . Under such alternatives, sometimes also known as contiguous alternatives, we have the following facts:

- $n^{1/2}\{\mu(\widehat{\alpha}_R, 0) - \mu(\alpha, \gamma)\}$  is asymptotically normally distributed but with nonzero mean, reflecting a bias when fitting the reduced model to full model (local) alternatives.
- The likelihood ratio test  $\mathcal{L}_n$  is still asymptotically noncentral chi-squared with noncentrality parameter, but now the noncentrality parameter converges to a constant.
- Thus, asymptotically BIC will now pick the reduced model with probability  $\rightarrow 1$ , and hence lead to biased estimates of parameters.