The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 553-566

Measurement Error Variance Estimation Based on Complex Survey Data with Subsample Re-Measurements

Sunyeong Heo[1] and John L. Eltinge[2]

Abstract

In many cases, the measurement error variances may be functions of the unknown true values or related covariates. This paper considers design-based estimators of the parameters of these variance functions based on the within-unit sample variances. This paper devotes to: (1) define an error scale factor δ; (2) develop estimators of the parameters of the linear measurement error variance function of the true values under large-sample and small-error conditions; (3) use propensity methods to adjust survey weights to account for possible selection effects at the replicate level. The proposed methods are applied to medical examination data from the U.S. Third *National Health and Nutrition Examination Survey* (NHANES III).

Keywords : Stratified multistage sampling design; Small-error approximation; Linear regression model with unequal variances; Logistic regression; U.S. Third National Health and Nutrition Examination Survey.

1. Introduction

A measurement error is generally defined to be the difference between an observed value and an underlying true value. Some authors, e.g., Grove (1991), refer to measurement errors as observed errors. If measurement errors are nontrivial, then estimators from classical methods may have corresponding nontrivial biases. For example, Fuller (1987, sec.1.1.) noted that if the predictor variables in a simple linear regression model are measured with error, the ordinary least squares estimators generally are biased. Also, the correlation between the dependent variable and the independent variables is generally reduced by the presence of measurement error.

1) Full-time Lecturer, Department of Statistics, Changwon National University, Changwon, 641-773, Korea.
E-mail : syheo@sarim.changwon.ac.kr
2) Senior Mathematical Statistician, U.S. Bureau of Labor Statistics, PSB 4915, 2 Massachusetts Avenue NE, Washington, DC 20212 U. S. A.

Since the 1940s, people have been concerned about various problems associated with measurement errors. See, e.g., Dalenius (1981) for a review of some early literature, and Biemer *et al.* (1991) for a more recent review.

Carroll and Stefanski (1990) defined a general form of a measurement error model. They considered three general approximate models for a response Y given Z when Z is a q -variate proxy for a p-variate predictor X $(q \geq p)$ and some of Z are measured with error. One of their important results is that when the measurement error is small, under additional conditions one can directly use Z in the place of X without accounting explicitly for the errors. They assumed that observations are stochastically independent. However, in a complex survey design using cluster sampling, observations are not independent within a cluster.

This paper will consider measurement error variance estimation for a known function under a finite population conditions and a specified complex sampling design. By using extensions of the Carroll and Stefanski results, we will derive estimators of the parameters when measurements errors are small.

In Section 2, we define a measurement error model and a measurement error variance function. In Section 3, we define a sampling design and an error scale factor δ, and develop estimators of the parameters defined in Section 2 under small error conditions. In Section 4, we apply the methods of this paper to data from the U.S. Third National Health and Nutrition Examination Survey (NHANES III).

2. Measurement Error Model and Measurement Error Variance Function

2.1 Measurement Error Model

Assume that we know only Z as a proxy for X and also assume that Z is unbiased for X. In addition, two replicate measurements are taken at each design point. Then following the notation of Carroll and Stefanski (1990), for a given x_t the model will be written as

$$Z_{tr} = x_t + \delta U_{tr} \tag{2.1}$$

for $t = 1, 2, \ldots, n$; $r = 1, 2$ where δ is a positive scale factor and the random variable U_{tr} has

$$E(U_{tr}|x_t) = 0 \text{ and } Var(U_{tr}|x_t) = \Omega(x_t, \gamma).$$

The $\Omega(x_t, \gamma)$ is a known function of parameter γ and x_t. In the following work we will denote $\Omega(x_t, \gamma)$ as Ω_t if it is not necessary to emphasize Ω_t being a function of (x_t, γ).

The random variable U_{tr} in model (2.1) can be written as $U_{tr} = \Omega_t^{1/2} d_{tr}$ where $d_{tr} = \Omega_t^{-1/2} U_{tr}$. The random variable d_{tr} is independent of x. In other words, U_{tr} depends

on x only through the scale factor $\Omega_t^{1/2}$. From this fact, we may write model (2.1) as

$$Z_{tr} = x_t + (\delta \Omega_t^{1/2}) d_{tr} \tag{2.2}$$

where the d_{tr} are independent and identically distributed with mean 0 and variance 1 for all t and r.

2.2 Measurement Error Variance Function

In many cases, the measurement error variances increase proportionally as the values of predictors increase or decrease. For a first order approximation, it sometimes suffices to model the measurement error variance $\Omega(x_t, \gamma)$ as a linear function of x, and this paper will focus on this approach.

The linear measurement error variance function of x will be written as

$$\Omega_t = \gamma_0 + \gamma_1 x_t \tag{2.3}$$

for a given x_t.

Davidian and Carroll (1987) discussed several methods of variance function estimation based on some transformations of absolute residuals from the current fit to the mean or sample standard deviations from replicates. Davidian (1990) compared efficiencies of different transformations based on sample standard deviations from replicates for contaminated normal distributions. Through simulation results, she showed that the square transformations may be more efficient than the log and the identity under normal distribution conditions.

In (2.1) we assumed that we do not know the values of x_t, and hence we can not calculate residuals from fitting of model (2.3). Therefore, we will consider the measurement error variance estimation based on the sample variances.

Under model (2.1), for a given value x_t and a known δ, an unbiased estimator of Ω_t is $\delta^{-2} S_t^2$ where $S_t^2 = (Z_{t1} - Z_{t2})^2/2 = \delta^2 S_{Ut}^2$ and $S_{Ut}^2 = (U_{t1} - U_{t2})^2/2$ is the sample variance within the tth unit. Note that $S_{Ut} = \delta^{-2} S_t^2$. In addition, an unbiased estimator of x_t is $\bar{Z}_{t\cdot} = \frac{1}{2}(Z_{t1} + Z_{t2})$.

3. Design Based Estimation of Measurement Error Variance

In this section, we will consider the estimation of $\gamma = (\gamma_0, \gamma_1)'$ in equation (2.3) based on data obtained from a complex sample selected from a finite population. When we have the true observations of entire population, γ will be a function of the population totals of the observations. The parameter γ can be estimated by a sample which is drawn by a specified

sampling method.

3.1 Sampling Design

We assume the following design condition which is quoted with minor modifications from the stratified multistage sampling design in Shao (1996, p.205-206).

(D.1) The population has been stratified into L strata with N_h clusters in the hth stratum. For the hth stratum, $n_h(\geq 2)$ clusters are selected independently across the strata. These first-stage clusters are selected with unequal selection probabilities and with replacement. Within the ith first-stage cluster in the hth stratum, $n_{hi} \geq 1$ ultimate units are sampled according to some sampling methods, $i=1,\cdots,n_h$, $h=1,\cdots,L$. The total number of ultimate units in the population is $N=\sum_{h=1}^{L}\sum_{i=1}^{N_h} N_{hi}$ and in the sample is $n=\sum_{h=1}^{L}\sum_{i=1}^{n_h} n_{hi}$. The total number of first-stage clusters in the sample is $n_F=\sum_{h=1}^{L} n_h$.

3.2 A Linear Measurement Error Variance Model

Under model (2.3) we may write a regression model such as

$$\Omega_{hij} = \gamma_0 + \gamma_1 x_{hij} + \varepsilon_{hij} \tag{3.1}$$

where ε_{hij} are independent and identically distributed with mean 0 and a finite variance σ_ε^2 for all (hij). The ε_{hij} account for the deviation of the Ω_{hij} from the line $\gamma_0+\gamma_1 x_{hij}$. Under model (2.1) when we have n observed values, $(Y_{\delta hij}, X_{\delta hij}) = (\delta^{-2}S^2_{hij}, \overline{Z}_{hij\cdot})$, for a given δ instead of (Ω_{hij}, x_{hij}), we can write a model,

$$\begin{aligned} y_{hij} &= \gamma_0 + \gamma_1 x_{hij} + \varepsilon_{hij} \\ (Y_{\delta hij}, X_{\delta hij}) &= (y_{hij}, x_{hij}) + (\zeta_{hij}, u_{hij}) \end{aligned} \tag{3.2}$$

where $y_{hij}=\Omega_{hij}$, $\zeta_{hij}=S^2_{Uhij}-\Omega_{hij}$ and $u_{hij}=\delta\,\overline{U}_{hij\cdot}$. The $\overline{U}_{hij\cdot}=\frac{1}{2}(U_{hij1}+U_{hij2})$. The variable ζ_{hij} is an independent $(0, \sigma_{\zeta\zeta hij})$ random variable with $\sigma_{\zeta\zeta hij}=\mathrm{Var}(S^2_{Uhij})$ and the variable u_{hij} is an independent $(0, \sigma_{uuhij})$ random variable with $\sigma_{uuhij}=\delta^2\Omega_{hij}/2$. Note that under model (2.2) $S^2_{Uhij}=\Omega_{hij}S^2_{dhij}$ where $S^2_{dhij}=2^{-1}(d_{hij1}-d_{hij2})^2$ and S^2_{dhij} is independent and identically distributed with mean 1 and a constant variance, say c. Therefore $\mathrm{Var}(S^2_{Uhij})=c\Omega^2_{hij}$. If we assume that the U_{hijr} follow a normal distribution, then $c=2$. We will assume here that the errors, ε_{hij}, in the regression equation (3.2) are independent of

$(x_{hij}, \zeta_{hij}, u_{hij})$ for all (hij).

For the following theory, we need to take expectations in two ways; one is based on the sampling design (D.1), the other is based on the model (3.2), especially for the error terms. We will use the notations $E_{D\xi}$ to denote expectation evaluated with respect to on both design and model, E_D based on only the sampling design, and E_ξ based on only the model.

For convenience, we will replace the triple subscript (hij) with the single subscript t from the following expressions if it is not necessary to specify strata, clusters and ultimate units.

3.3 Estimation of moments of the finite population

Under the survey design (D.1) and model (3.2), when we have N true values, (y_t, x_t), define

$$\Sigma_{xx} = N^{-1} \sum_{t=1}^{N} (1\ x_t)'(1\ x_t) \text{ and } \Sigma_{xy} = N^{-1} \sum_{t=1}^{N} (1\ x_t)' y_t$$

and when we have N observations of $(Y_{\delta t}, X_{\delta t})$ instead of (y_t, x_t), define

$$\Sigma_{X_\delta X_\delta} = N^{-1} \sum_{t=1}^{N} (1\ X_{\delta t})'(1\ X_{\delta t}) \text{ and } \Sigma_{X_\delta Y_\delta} = N^{-1} \sum_{t=1}^{N} (1\ X_{\delta t})' Y_{\delta t}$$

and also define

$$\Sigma_{uu} = \text{diag}(0, \sigma_{uu}) \text{ and } \Sigma_{\delta uu} = \text{diag}(0, \sigma_{\delta uu})$$

with $\sigma_{uu} = N^{-1} \sum_{t=1}^{N} \sigma_{uut}$ and $\sigma_{\delta uu} = N^{-1} \sum_{t=1}^{N} \hat{\sigma}_{uut}$ where $\sigma_{uut} = \delta^2 \Omega_t / 2$ and $\hat{\sigma}_{uut} = S_t^2 / 2$.

When we have n sampled true values, (y_t, x_t), from a finite population, Σ_{xx} and Σ_{xy} can be estimated by

$$M_{xx} = N^{-1} \sum_{t=1}^{n} w_t (1\ x_t)'(1\ x_t) \text{ and } M_{xy} = N^{-1} \sum_{t=1}^{n} w_t (1\ x_t)' y_t$$

where w_t is a unit-level survey weight. In addition, when we have observations of $(Y_{\delta t}, X_{\delta t})$ in the place of (y_t, x_t), we may have

$$M_{X_\delta X_\delta} = N^{-1} \sum_{t=1}^{n} w_t (1\ X_{\delta t})'(1\ X_{\delta t}) \text{ and } M_{X_\delta Y_\delta} = N^{-1} \sum_{t=1}^{n} w_t (1\ X_{\delta t})' Y_{\delta t}$$

instead of M_{xx} and M_{xy}. In addition, Σ_{uu} can be estimated by $S_{\delta uu} = \text{diag}(0, \hat{\sigma}_{\delta uu})$ where $\hat{\sigma}_{\delta uu} = N^{-1} \sum_{t=1}^{n} w_t \hat{\sigma}_{uut}$.

3.4 Definition and Estimation of δ

To define the error scale factor δ in (2.1), the concept of a sequence of finite populations is required. By quoting Shao's notations, we will consider a sequence of finite populations

indexed by $k=1,2,\ldots$ with population size N_k for each k. Then the population quantities L, N, N_h, N_{hi}, the sample sizes n, n_h, n_{hi}, the sample values y_{hij} and the survey weight w_{hij} depend on the index k (Shao, 1996, p.210).

Definition 3.1. Use model (3.2), define

$$\delta^2 = q^{-1}\,\mathrm{tr}\left[\Sigma_{xx}^{-1/2}\{\mathrm{diag}\,(0,\Omega)\}\Sigma_{xx}^{-1/2}\right]$$

where $\Omega = plim\left\{N_k^{-1}\sum_{t\in U_k}\mathrm{Var}(\bar{Z}_{t\cdot}-x_t)\right\}$, $\Sigma_{xx}=plim\left\{N_k^{-1}\sum_{t\in U_k}(1,x_t)'(1,x_t)\right\}$, U_k is the kth full finite population, and q is the rank of the matrix $\left[\Sigma_{xx}^{-1/2}\{\mathrm{diag}(0,\Omega)\}\Sigma_{xx}^{-1/2}\right]$.

In the Definition 3.1, the notation $plim(X_k)=X$ denotes that the sequence of a random variable X_k converges in probability to a limit X. The $tr(A)$ denotes the trace of matrix A.

To motivate and illustrate Definition 3.1 of δ^2, note that $M_{X_\delta X_\delta}$ can be expressed such as

$$M_{X_\delta X_\delta} = M_{xx} + S_{\delta uu} + R \tag{3.3}$$

where R denotes the remainder.

Under model (2.2), design (D.1), and additional regularity conditions, R is at most order in probability $n_F^{-1/2}\delta$ and $\hat{\sigma}_{\delta uu}-\hat{\sigma}_{uu}=O_p(n_F^{-1/2}\delta^2)$ where $\hat{\sigma}_{uu}=N^{-1}\sum_{t=1}^{n}w_t\sigma_{uut}$. For more details of the order of R, see the Appendix. When we assume that M_{xx} is nonsingular,

$$M_{xx}^{-1}M_{X_\delta X_\delta} = 1 + M_{xx}^{-1}S_{\delta uu} + M_{xx}^{-1}R.$$

Under design (D.1), model (2.2), and additional regularity conditions,

$$M_{xx}^{-1}R = O(1)\cdot O_p(n_F^{-1/2}\delta) = O_p(n_F^{-1/2}\delta)$$

and

$$M_{xx}^{-1}S_{\delta uu} = M_{xx}^{-1}S_{uu} + M_{xx}^{-1}(S_{\delta uu}-S_{uu}) = O(\delta^2) + O(1)\cdot O_p(n_F^{-1/2}\delta^2)$$

where $S_{uu}=\mathrm{diag}(0,\ \hat{\sigma}_{uu})$.

Consequently, the performance of $M_{X_\delta X_\delta}$ will depend primarily on $M_{xx}^{-1}S_{\delta uu}$ than $M_{xx}^{-1}R$ provided we have both δ small and n_F large. Thus, the size of $M_{X_\delta X_\delta}$ relative to M_{xx} depends on the size of $M_{xx}^{-1}S_{\delta uu}$.

Note that the trace of $(M_{xx}^{-1/2}S_{\delta uu}M_{xx}^{-1/2})$ is equal to the trace of $M_{xx}^{-1}S_{\delta uu}$. Also, note that under model (3.2), $E_\xi(\hat{\sigma}_{\delta uu}|x)=\hat{\sigma}_{uu}$ and $E_\xi(\hat{M}_{xx}|x)=M_{xx}$ where $\hat{M}_{xx}=M_{X_\delta X_\delta}-S_{\delta uu}$. Thus, δ can be estimated by

$$\hat{\delta}=\{q^{-1}tr(\ \hat{M}_{xx}^{-1/2}S_{\delta uu}\ \hat{M}_{xx}^{-1/2})\}^{1/2} \tag{3.4}$$

3.5 Measurement Error Variance Estimation under Small Error Approximation

When the measurement error is small, by the result of Carroll and Stefanski (1990, p. 654), we can rewrite the model (3.2) such as

$$Y_{\delta t}=\gamma_0+\gamma_1 X_{\delta t}+\psi_t \tag{3.5}$$

where $\psi_t=(\varepsilon_t+\zeta_t)-\gamma_1 u_t$ with

$$E_\xi(\psi_t|x_t)=0 \text{ and } V_\xi(\psi_t|x_t)=\sigma_\varepsilon^2+c\Omega_t^2+\gamma_1^2(\delta^2\Omega_t/2).$$

The variance of ψ_t for given value x_t, $Var_\xi(\psi_t|x_t)$, goes to $\sigma_\varepsilon^2+c\Omega_t^2$ as $\delta\to 0$. The model (3.5) has the form of a simple linear regression model with unequal variances. Therefore, under the sampling design (D.1) we have an ordinary least squares estimator of γ,

$$\hat{\gamma}=M_{X_sX_s}^{-1}M_{X_sY_s} \tag{3.6}$$

where $\hat{\gamma}=(\hat{\gamma}_0,\hat{\gamma}_1)'$. Under the design (D.1), model (3.5) and additional regularity conditions, $M_{X_sX_s}-\Sigma_{X_sX_s}=O_p(n_F^{-1/2})$ and $M_{X_sY_s}-\Sigma_{X_sY_s}=O_p(n_F^{-1/2})$. These results and additional routine arguments show that $\hat{\gamma}-\gamma=O_p(n_F^{-1/2})$.

4. Application to the U.S. NHANES III Data

4.1 The U.S. NHANES III Data

The U.S. Third National Health and Nutrition Examination Survey (NHANES III) was conducted for the U.S. National Center for Health Statistics (NCHS) to assess the health and nutritional status of the non-institutionalized civilian population in the United States. NHANES III is a large-scale sample survey based on a stratified multistage design with 49 strata. Within each stratum, two primary sample units (PSUs, roughly equivalent to counties) are selected with unequal probabilities. Additional levels of sampling select secondary units (roughly equivalent to city blocks), households and individual persons. Each selected person is asked to complete a questionnaire and to participate in a very thorough medical examination.

As part of NHANES III, the NCHS considered using a formal two-phase sampling design to obtain replicate measurements from a relatively small subset of the group of original respondents. However, this ultimately was not feasible due to scheduling constraints for the medical examination equipment and other factors. As an alternative, participants in a first interview and examination were asked at the end of the examination whether they would be willing to participate in a second interview and examination. Those who agreed were listed

and were selected for re-examination as scheduling permitted. An additional complication that arose with bone mineral density (BMD) measurements was that some sampled persons did not provide measurements even though they participated in the remainder of the interview and examination. For example, any woman who indicated she might be pregnant was excluded from the BMD measurements.

In this research, we restricted the analysis to adults aged 20 and up; very few replicate measurements were collected from children. The measurement we are interested in is total region bone mineral density (TOBMD) measurement. Two TOBMD measurements were obtained only 1,108 persons among 16,573 adults.

4.2 Survey Weight Adjustment

Since replicate measurements were obtained from only a small subset of the original respondents, the survey weights, say w_{1hij}, need to be adjusted to account for possible selection effects at the replication level. If we know the probability, say p_{hij}, that a unit in the original respondents' group given sample gives replicate measurements, then the adjusted weight can be expressed by

$$w_{2hij} = w_{1hij} / p_{hij} \tag{3.7}$$

and the population size, N, is equal to $\sum_{hij \in s_r} w_{2hij}$ where s_r represents the set of all units (hij) responding at both stages. Thus, by using this adjusted weights, w_{2hij}, to $M_{X_\delta X_\delta}$, $M_{X_\delta Y_\delta}$, $S_{\delta uu}$, we obtain the same results as section 3 for the measurement error variance estimator.

The process of which a unit responds at the second stage can be modeled as a Bernoulli(p) random variable. The probability p can be considered a function, say $p = p(x)$, of some auxiliary variables x that are observed for both respondents and nonrespondents.

Define the response indicator $r_{hij} = 1$ if a unit (hij) gives replicate measurement; 0 other. The probability, p_{hij}, can be estimated by a logistic regression model such as

$$\log[p(x)/\{1-p(x)\}] = x'\beta \tag{3.8}$$

(Eltinge, Heo, and Lee, 1997).

To find a model that explains well the probability that one gives two TOBMD measurements, we use the logistic regression model in (3.8). We anticipate that a respondents's race/ethnic origin, gender, age and resident place will affect the probability; specific explanatory indicator variables are reported in Table 4.1. Exploratory analysis led to final model coefficient estimates reported in Table 4.2. Using the final model we estimate the probability, call it $\hat{p}_{hij}$. Replacing p_{hij} with $\hat{p}_{hij}$ in (3.7) we adjust the original weights.

Table 4.3 shows the summary of survey weights. The total survey examination weights of

the original NHANES III is 1.772×10^8. The total weights attached to persons giving two TOBMD measurements are only 12,976,478 which is only 7.32% of entire population size. After adjusting for missing values on replicate TOBMD measurements, the total weights for

Table 4.1 Explanatory indicator variables for the logistic regression model.

Variable Name	Group Indicated
(Baseline Region)	(Northeast)
Other Region	Midwest, South, and West
(Baseline Race/Ethnic)	(Non-Hispanic White)
Black	Non-Hispanic Black
MAmer	Mexican-American
Other	Other
ORegion*Other	Midwest, South and West × Other
(Baseline Gender)	(Male)
Female	Female
(Baseline Age)	(70+)
Age20	20-29
Age30	30-39
Age40	40-49
Age50	50-59
Age60	60-69
Age *i*F	Age i × Female, i = 20,30,40,50,60

Table 4.2 Logistic regression coefficient point estimates, standard errors, approximate 95% confidence intervals for the second-phase TOBMD sample selection propensity model.

Predictor	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	$(\hat{\beta}_{iL}, \hat{\beta}_{iU})$
Intercept	-2.433	0.145	(-2.724, -2.141)
Other Region	-0.110	0.121	(-0.354, 0.133)
Black	-0.183	0.089	(-0.362, -0.003)
MAmer	-0.291	0.103	(-0.499, -0.083)
Other	0.269	0.267	(-0.267, 0.805)
ORegion*Other	-1.251	0.451	(-2.158, -0.344)
Female	-0.412	0.168	(-0.749, -0.075)
Age20	-0.019	0.208	(-0.437, 0.398)
Age30	-0.258	0.210	(-0.680, 0.164)
Age40	-0.015	0.240	(-0.498, 0.468)
Age50	0.279	0.246	(-0.215, 0.773)
Age60	0.611	0.144	(0.322, 0.900)
Age20F	0.183	0.249	(-0.318, 0.684)
Age30F	0.612	0.219	(0.172, 1.052)
Age40F	0.669	0.223	(0.221, 1.118)
Age50F	0.447	0.268	(-0.092, 0.986)
Age60F	0.105	0.259	(-0.415, 0.626)

Table 4.3 The summary of survey weights for the original NHANES III design.

Weight		Total
w_{1hij}	entire sample	1.772×10^8
	replicate only	12,976,478
w_{2hij}	replicate only	1.771×10^8

Table 4.4 Weighted regression coefficients estimates and standard errors and approximate 95% confidence intervals for the measurement error variance regression model.

Predictor	$\hat{\gamma}_i \times 10^4$	$se(\hat{\gamma}_i) \times 10^4$	$(\hat{\gamma}_{iL}, \hat{\gamma}_{iU}) \times 10^4$
Intercept	-0.995	4.017	(-9.069, 7.078)
$\bar{Z}_{t\cdot}$	13.644	4.717	(4.165, 23.123)
Age20	-7.558	2.390	(-12.362, -2.754)
Age30	-4.151	1.716	(-7.598, -0.703)
Age40	-4.082	2.237	(-8.578, 0.413)
Age50	-4.807	2.077	(-8.981, -0.634)

persons to give two TOBMD measurements is 1.771×10^8. Therefore, the estimated population size, i.e., total weights, is almost equal to the true population size. Therefore, for the following calculations we will use w_{2hij}.

4.3 Estimate of Error Scale Factor

From expression (3.4), $\hat{\delta}$ of total BMD measurements is 0.0653. By experience of empirical analysis, we may consider $\hat{\delta} = 0.0653$ as small. Therefore, for the following measurement error variance model selection, we ignore misclassification errors. We use the model (3.5) for estimation of measurement error variance.

4.4 Measurement Error Variance Model Selection

For the following model fitting, the dependent variable is squared differences between two total BMD measurements, $(Z_{t1} - Z_{t2})^2$, not within sample variance $S_t^2 = (Z_{t1} - Z_{t2})^2/2$. Since, the constant 2^{-1} does not effect on the test statistics or p-value for test $H_0: \gamma_i = 0$. Dividing $(Z_{t1} - Z_{t2})^2$ by 2 causes the coefficient estimates and its standard errors to reduce by half and the confidence interval too. We consider the same four demographic variables as propensity model selection as explanatory variables. Exploratory analysis led to the final model coefficient estimated reported in Table 4.4. The model giving Table 4.4 can be written as the following

$$\Omega_t = \gamma_0 + \gamma_1 x_t + \gamma_{21}\text{Age20} + \gamma_{22}\text{Age30} + \gamma_{23}\text{Age40} + \gamma_{24}\text{Age50}. \quad (4.1)$$

In model (4.1), the persons aged 60's and up are base group for age.

Table 4.4 shows that the measurement error variances increase as the $\bar{Z}_t$. increase. For age variable, $Age20$, $Age30$, $Age50$ are significant at $\alpha = 0.05$. Specially the persons aged 20's have smallest measurement error variance.

This estimated measurement error variances can be used for estimation of regression coefficients, quantiles or disease-prevalence rates that may be seriously influenced by measurement error.

5. Conclusion

5.1 Summary of Results

We have used the assumptions that the observed values are unbiased for the corresponding true values and that measurement error variance is a linear function of true values. Under

these assumptions and additional conditions, the ordinary least squared estimator of the measurement error variance function parameter, $\hat{\gamma}$, is consistent provided the error scale factor is small. This result is an extension of previous work by Carroll and Stefanski (1990) on model based errors-in-variables estimation under small error approximations.

Next, we illustrated some of the proposed estimation methods with an analysis of bone mineral density measurements from the U.S. Third National Health and Nutrition Examination Survey (NHANES III). Detailed examination of model fitting results indicated that the measurement error variance is a function of true value and age.

The original survey weight need to be adjusted to account for nonuniform mechanisms in the selection of replicate-measurement subsample. In section 3, we considered the propensity models for adjustment of original survey weights to account the probability that a given sampled unit provided replicate measurement. We applied the proposed methods to develop a propensity model for selection of replicate measurement subsamples in the NHANES III; and to construct an associate weighting adjustments.

5.2 Future Research

The present modeling work can be extended in several possible directions. For example, we assumed here that the measurement error variance is a linear function of the true value. However, the method developed here can be extended to a non-linear function of true value under a complex sampling design.

Appendix

Lemma A. Assume model (2.2) and the sampling design (D.1). Then, with additional regularity conditions the remainder R in (3.3) is at most of order in probability $n_F^{-1/2}\delta$ for $\delta \leq 1$.

Proof From the definition of moments in section 3.3, we can write such as

$$M_{X_\delta X_\delta} = M_{xx} + S_{\delta uu} + R$$

where R is a 2×2 matrix with elements $R_{11} = 0$, $R_{12} = R_{21} = \delta N^{-1} \sum_{t=1}^{n} w_t(\Omega_t^{1/2}\ \bar{d}_{t\cdot})$ and $R_{22} = \delta N^{-1} \sum_{t=1}^{n} w_t(2x_t\Omega_t^{1/2}\ \bar{d}_{t\cdot}) + \delta^2 N^{-1} \sum_{t=1}^{n} w_t \Omega_t d_{t1} d_{t2}$. To show $R = O_p(n_F^{-1/2}\delta)$, it is enough to examine each element of matrix R is at most of order in probability $n_F^{-1/2}\delta$.

The R_{12} can be expressed by $R_{12} = \delta A_{12}$ where $A_{12} = N^{-1} \sum_{t=1}^{n} w_t(\Omega_t^{1/2}\ \bar{d}_{t\cdot})$. The A_{12} is

a survey mean of $\Omega_{1/2}\,\bar{d}_{t\cdot}$. Then, we need to show $A_{12} = O_p(n_F^{-1/2})$. Under design (D.1) and model (3.2), the expectations of A_{12} is $E_{D\xi} = E_D\left\{N^{-1}\sum_{t=1}^{n} w_t \Omega_t^{1/2} E_\xi(\bar{d}_{t\cdot})\right\} = 0$ since $E_\xi(\bar{d}_{t\cdot}) = 0$. Hence, with additional regularity conditions $A_{12} = O_p(n_F^{-1/2})$ (Shao, p. 210-211). Since $R_{12} = \delta A_{12}$, $R_{12} = O_p(n_F^{-1/2}\delta)$

Now, let us express R_{22} as $R_{22} = \delta B_{22} + \delta^2 C_{22}$ where $B_{22} = N^{-1}\sum_{t=1}^{n} w_t(2x_t\Omega_t^{1/2}\,\bar{d}_{t\cdot})$ and $C_{22} = N^{-1}\sum_{t=1}^{n} w_t \Omega_t d_{t1} d_{t2}$. By the same discussions as A_{12}, $B_{22} = O_p(n_F^{-1/2})$ and $\delta B_{22} = O_p(n_F^{-1/2}\delta)$.

Under design (D.1) and model (2.2), the expectation of C_{22} is $E_{D\xi}(C_{22}) = E_D\left\{N^{-1}\sum_{t=1}^{n} w_t \Omega_t E_\xi(d_{t1} d_{t2})\right\} = 0$ since $E_\xi(d_{t1} d_{t2}) = E_\xi(d_{t1})E_\xi(d_{t2}) = 0$. By the similar discussions as A_{12}, $C_{22} = O_p(n_F^{-1/2})$ and $\delta^2 C_{22} = O_p(n_F^{-1/2}\delta^2)$.

Since $R_{22} = \delta B_{22} + \delta^2 C_{22}$, $R_{22} = O_p(n_F^{-1/2}\delta) + O_p(n_F^{-1/2}\delta^2) = O_p(\max\{n_F^{-1/2}\delta,\ n_F^{-1/2}\delta^2\})$. Thus $R_{22} = O_p(n_F^{-1/2}\delta)$ for $\delta \le 1$ and $R_{22} = O_p(n_F^{-1/2}\delta^2)$ for $\delta > 1$.

For large n_F, the effect of the size of δ on the convergence rate is relatively small. Hence, from the main body of text we considered that the order of R is $R = O_p(n_F^{-1/2}\delta)$.

Acknowledgements

Authors thank D. Brody, M.A. Carroll, A. Looker, R. Murphy, and V.L. Parsons of the U.S. National Center for Health Statistics (NCHS) for providing the NHANES III data used in this paper. and for helpful comments on the NHANES III data. The views expressed in this paper are those of authors and do not necessarily represent the policies of NCHS.

Reference

[1] Biemer P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (eds.). (1991). *Measurement Errors in Surveys.* John Wiley & Sons, New York.

[2] Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association,* Vol. 85, 652-663.

[3] Dalenius, T. E. (1981). The survey statistician's responsibility for both sampling and

measurement error. In D. Krewski, R. Platex, and J. N. K. Rao (eds.), *Current Topics in Survey Sampling*, 17-29. Academic Press, New York.

[4] Davidian, M. (1990). Estimation of variance functions in assays with possible unequal replication and nonnormal data. *Biometrika*, Vol. 77, 43-54.

[5] Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, Vol. 82, 1079-1091.

[6] Eltinge, J. L., Heo, S., and Lee, S. R. (1997). use of propensity methods in the analysis of subsample re-measurements for NHANES III. In *Proceedings of the Survey Methods Section, Statistical Scociety of Canada*, 27-36.

[7] Fuller, W. A. (1987). *Measurement Error Models*. Jonh Wiley, New York.

[8] Grove, R. M. (1991). Measurement error across the disciplines. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, 1-25. John Wiley & Sons, New York.

[9] Shao, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics,* Vol. 27, 203-254.

[Received April 2003, Accepted July 2003]