J. of the Chosun Natural Science
Vol. 5, No. 3 (2012) pp. 157 – 162

Linear Measurement Error Variance Estimation based on the Complex Sample Survey Data

Sunyeong Heo† and Duk-Joon Chang

Abstract

Measurement error is one of main source of error in survey. It is generally defined as the difference between an observed value and an underlying true value. An observed value with error may be expressed as a function of the true value plus error term. In some cases, the measurement error variance may be also a function of the unknown true value. The error variance function can be rewritten as a function of true value multiplied by a scale factor. This research explore methods for estimation of the measurement error variance based on the data from complex sampling design. We consider the case in which the variance of mesurement error is a linear function of unknown true value, and the error variance scale factor is small. We applied our results to the U.S. Third National Health and Nutrition Examination Survey (the U.S. NHANES Ⅲ) data for empirical analyses, which has replicate measurements for relatively small subset of initial respondents's group.

Key words : Complex Sampling Design, Measurement Error Variance, Logistic Regression

1. Introduction

Errors in surveys arise two general sources; sampling error and nonsampling error. The major sources of nonsampling error generally divided further into three categories; coverage, nonresponse and measurement error[1]. Measurement error is generally defined as the difference between an observed value and an underlying true value.

Since the 1940s, people have been concerned about various problems associated with measurement errors. See, e.g., Dalenius[2] for a review of some early literature, and Biemer *et al.*[3] for a more recent review.

The purpose of this research is to develop design-based estimators of the parameters of measurement error variance functions based on data from complex survey. This problem is of practical interest because in many practical analyses measurement error variances are more realistically viewed as functions of the true values. In some cases, measurement error variances increase in proportion to the values of predictors.

Davidian and Carroll[4] discussed several methods of variance function estimation under a heteroscedastic regression model. Their simulation results showed that when data are approximately normally distributed and the mean response function is known, using absolute residuals or squared residuals as responses for estimation of heteroscedastic variance functions gives better efficiency than using sample standard deviations or sample variances from the small numbers of replicates (less than ten replicates) at each design point. However, in many cases the exact form of the mean response function is generally unknown, and the residuals from an incorrect mean response function may cause a wrong estimate of variance function. Therefore, when we do not know its exact form, it may give better estimates of variance functions to use sample standard deviations or sample variances from replicates as responses than absolute residuals or squared residuals.

Carroll and Cline[5] discussed estimation of the coefficients of a heteroscedastic linear regression model. They developed a related asymptotic theory in which the inverse of the sample variances is used as estimated weights to estimate the regression coefficients. They find that using sample variances for estimating weights

Department of Statistics, Changwon National University Changwon 641-773, Korea

†Corresponding author : syheo@changwon.ac.kr

(Received : June 8, 2012, Revised : June 20, 2012, Accepted : June 24, 2012)

is inefficient unless the number of replicates are large, e.g. at least ten.

Davidian[6] compared efficiencies of different transformations based on sample standard deviations. She presented the asymptotic relative efficiencies obtained from the different number of replicates from contaminated normal distributions with different levels of contamination proportions. Through simulation results, she showed that as the contamination proportion increases, the identity transformation was more efficient than the square transformation. In addition, the identity transformation may be more efficient than the log transformation especially for small numbers of replicates (two or three replicates). However, the square transformation may be more efficient than the log and the identity under normal distribution conditions. In the present research we will use the square transformation of sample standard deviation.

Following the notation of Carroll and Stefanski[7], a general form of a measurement error model is defined by

$$W = c(X, \eta) + \delta U$$

with $E(U|X=x)=0$ and $Cov(U|X=x)=\Omega(\mathrm{x}, \eta, \gamma)$ where X is a p-variate predictor, W is a q-variate ($q \geq p$) proxy for X, and $c(\cdot, \cdot)$ and $\Omega(\cdot, \cdot)$ are known functions of the column vector of parameters $\Lambda = (\eta^t, \gamma^t, \delta^t)^t$. In certain models some components of Λ may be known. They considered three general approximate models for a response Y given W when some of the predictors X are measured with error. One of their important results is that when the measurement error is small, under additional conditions W can be directly used in the place of X without accounting explicitly for the errors.

Carroll and Stefanski assumed that observations are stochastically independent. However, in a complex survey design using cluster sampling, observations are not independent within a cluster. Fuller[8] gave a complex survey sample based errors-in-variables estimator of a vector of regression coefficients. He assumed measurement error vectors are independent and identically distributed at the super-population level. Under this assumption and some additional regularity conditions, he established the consistency and asymptotic normality of his errors-in-variables estimator.

In the present research, we consider the sample variances as responses and combine Carroll and Stefanski and Fuller to estimate measurement error variance based on data from complex sample design under the assumption of small measurement error.

2. Experimental

2.1 Measurement Error Models

Assume that W is unbiased for X, i.e., $c(x, \eta) = x$, and also assume that only W is known but not x. In addition, two replicate measurements are taken at each design point. Then, for a given x_t, the model is written as

$$W_{tr} = x_t + \delta U_{tr} \qquad (2.1)$$

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

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

In the following work, Ω_t is used to denote $\Omega_t(x_t, \gamma)$ if it is not necessary to emphasize that Ω is a function of (x_t, γ).

Now, let the random variable U_{tr} in (2.1) denote as $U_{tr} = \Omega_t^{1/2} d_{tr}$, and assume that d_{tr} is independent of x, so that U_{tr} depends on x only through the scale factor $\Omega_t^{1/2}$. From this setting, the model (2.1) may be written as

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

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

2.2 Linear Measurement Error Variance Function

In many cases, the measurement error variances increase proportionally as the values of predictors increase or decrease. In the case the measurement error variance $\Omega(x_t, \gamma)$ can be modeled as a linear function of x_t as the following

$$\Omega_t = \gamma_0 + \gamma_1 x_t \qquad (2.3)$$

for a given x_t. 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 $S_t^2 = (W_{t1} - W_{t2})^2 / 2 = \delta^2 S_{Ut}^2$ and $S_{Ut}^2 = 1(U_{t1} - U_{t2})^{2/2}$ is the sample variance within the t-th unit.

2.3 Design Based Estimation of Measurement Error Variance

Under finite population, if one has 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. Here the following sampling design is assumed, which is slightly modified from the stratified multistage sampling design in Shao[9]: The population has been stratified into L strata with N_h clusters in the hth stratum. From the hth stratum, $n_h = 2$ clusters are selected independently across the strata. These first-stage clusters are selected by using unequal probability p_{hi} without replacement. Within the ith first-stage cluster in the hth stratum, n_{hi} ultimate units are sampled with selection probability p_{hij} from N_{hi} units according to some sampling methods, $j = 1,...,n_{hi}$, $i = 1,2$, $h = 1,..., L$. In addition, it is assumed that for each h, N_h is large relative to $n_h = 2$, and so the sample can be treated as if the first-stage clusters are drawn with replacement. Under the above sampling design, the population total for observation y is estimated by

$$\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{ni}} w_{hij} y_{hij} .$$

where w_{hij} is the suvey weight associate element (hij). From the following expression, the triple subscript (hij) is replaced with the single subscript t if it is not necessary to specify strata, clusters, and ultimate units.

Under the model (2.3), a regression model may be written such as

$$\Omega_t = \gamma_0 + \gamma_1 x_t + q_t \qquad (2.4)$$

where q_t are independent and identically distributed with mean 0 and variance σ^2 for all t. The q_t account for the deviation of Ω_t from the line $\gamma_0 + \gamma_1 x_t$. Under model (2.1), n values $(Y_{\delta t}, X_{\delta t}) = (\delta^2 S_t^2, \overline{W}_t)$ for a given δ are observed in the place of (Ω_t, x_t), and the model (2.4) may be expressed such as

$$y_t = \gamma_0 + \gamma_1 x_t + q_t \text{ and } \begin{pmatrix} Y_{\delta t} \\ X_{\delta t} \end{pmatrix} = \begin{pmatrix} y_t \\ x_t \end{pmatrix} + \begin{pmatrix} w_t \\ u_t \end{pmatrix} \qquad (2.5)$$

where $y_t = \Omega_t$, $\omega_t = S_{Ut}^2 - \Omega_t$ and $u_t = \delta \overline{U}_t$. The variable ω_t is independent $(0, \omega_{\omega\omega t})$ random variable with $\sigma_{\omega\omega t} = Var(S_{Ut}^2)$, and the variable u_t is an independent $(0, \sigma_{uut})$ random variable with $\sigma_{uut} = \delta^2 \Omega_t / 2$. Note that under model (2.2) $S_{Ut}^2 = \Omega_t S_{dt}^2$ where $S_{dt}^2 = 2^{-1}(d_{t1} - d_{t2})^2$ and S_{dt}^2 is independent and identically distributed with mean 1 and a constant variance, say c. Now, let here assume that the errors q_t in the regression equation are independent of (x_t, ω_t, u_t) for all t.

Under the above mention sampling design and model (2.5), define

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

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

for N observations of $(Y_{\delta t}, X_{\delta t})$ and also define

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

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

for n sampled observations.

Now when the measurement error is small, by the result of Carroll and Stefanski (1990) the model (2.5) can be rewritten such as

$$Y_{\delta t} = \gamma_0 + X_{\delta t} \gamma_1 + \kappa_t \qquad (2.6)$$

where $\kappa_t = (q_t + \omega_t) - u_y \gamma_1$ with $E_\xi(\kappa_t | x_t) = 0$ and $Var_\xi(\kappa_t | x_t) = \sigma^2 + c\Omega_t^2 + \gamma_1(\delta^2 \Omega_t / 2)$. The variance of κ_t for given value x_t goes to $Var_\xi(\kappa_t | x_t) = \sigma^2 + c\Omega_t^2 + \gamma_1(\delta^2 \Omega_t / 2)$ as $\delta \rightarrow 0$, and the model (2.6) has the form of a simple linear regression model with unequal variances. Therefore, the above sampling design, the ordinary least squares estimator of γ is

$$\hat{\gamma} = M_{X_\delta X_\delta}^{-1} M_{X_\delta Y_\delta} \qquad (2.7)$$

where $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1)'$ Under some regularity conditions, $M_{X_\delta X_\delta} - \Sigma_{X_\delta X_\delta} \xrightarrow{p} 0$ and $M_{X_\delta Y_\delta} - \Sigma_{X_\delta Y_\delta} \xrightarrow{p} 0$ [9]. This result and some additional routine arguments show that $\hat{\gamma} - \gamma = O_p(n^{-1/2})$.

2.4 Propensity Model

Form the above sections, we considered that all units in a sample give replicates. However in many survey applications it is difficult or impossible to obtain replicate measurements from all sample units. In such applications, the survey weights, w_{1hij}, may need to be adjusted to account for possible selection effects at the replicate level. Specifically if we know the probability, say π_{hij}, that a unit in a given sample gives replicate measurements, then the adjusted weight can be expressed by

$$w_{2hij} = w_{1hij} / \pi_{hij}$$

and the population size, N, is equal to $\sum_{hij \in s_r} w_{2hij}$ where s_r represents the set of all units in the sample giving replicates. Thus, by applying this adjusted weight, w_{2hij}, for $(M_{X_\delta X_\delta}, M_{X_\delta Y_\delta})$, the estimator of γ in (2.6) and the same asymptotic result are obtained.

In practical settings, one generally does not know the true probability, π_{hij}. However, the process of which a unit gives a replicate can be modeled as a Bernoulli(π) random variable. The probability π can be considered as a function of some auxiliary variables x that are observed for both respondents and nonrespondents[10]. Define the response indicator $r_{hij} = 1$ if a unit (hij) gives replicate measurement; 0 other. Thus, the model is that the r_{hij} are independent Bernoulli (π_{hij}) random variables where π_{hij} is a probability conditional on some auxiliary variables, x, and conditional on the inclusion of person (hij) in the original sample. The probability, π_{hij}, can be estimated by a logistic regression model such as

$$\log[\pi(x) / \{1 - \pi(x)|] = x'\beta.$$

3. Results and Discussion

The above proposed idea is applied to the U.S. Third National Health and Nutrition Examination Survey (NHANES III). The U.S. NHANES III is a large-scale sample survey based stratified multistage design with 49 strata. Within each stratum, two primary sample units were selected with unequal probabilities. Additional levels of sampling selected households and individual persons. For empirical analysis here, only measurements from adults aged 20 and up are considered because very few replicate measurements were collected from children.

The U.S NHANES III data encountered measurement errors in analyses of bone mineral density (BMD) measurements, and a relatively small number of replicate BMD measurements was taken from the set of initial respondents. The number of initial adult respondents for BMD measurements was 14,646, and only 1,108 persons out of 14,646 gave replicate measurements.

It is necessary first to estimate the probability that one gives two BMD measurements, and a logistic regression model is used for it. The U.S NHANES•ð measured four BMD measurements and all four have the same numbers of initial respondents and replicates. Therefore, in selecting a model for the estimation of the probability, it is enough to derive models for only one BMD measurement. In this empirical analysis, total bone mineral density measurement is used for logistic model selection and linear measurement error variance estimation.

It is anticipated that a respondent's race/ethnic origin, gender, age and resident place will affect the probability of providing two measurements. After several fittings of logistic regression model for different set of main effects and interaction effects of above selected explanatory variables, a final model was selected. Table 1 shows the explanations of the explanatory indicator variables used for the final model. Table 2 shows logistic regression coefficient estimates and its standard errors and approximate 95% confidence intervals for the final

Table 1. Explanatory indicator variables for the final logistic regression model

Predictor	Group Indicator
(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,West×Other
(Baseline Gender)	(Male)
Female	Female
BlackF	Black×Female
MAmerF	Mexican-American×Female
OtherF	Other×Female
(Baseline Age)	(70+)
Age20	20-29
Age30	30-39
Age40	40-49
Age50	50-59
Age60	60-69

model. Using the final model, the probability of giving two measurements, call it $\hat{p}_{hij}$, and new weights are defined by $w_{2hij}=w_{1hij}/\hat{p}_{hij}$ where w_{1hij} is the original survey weight. Table 3 shows the summary of survey weights.

Now, using the new weight w_{2hij}, measurement error variance is estimated by the model (2.6) for total BMD measurements. The dependent variable is sample variances $S_t^2=(W_{t1}-W_{t2})^2/2$. Fig. 1 shows the scatter plot of sample means $\overline{W}_t=(W_{t1}+W_{t2})/2$ versus sample variances of two BMD measurements. The plot shows that as sample means increase, sample variances have a increasing trend too. Table 4 shows the estimates of survey weighted regression model fitting of S_t^2 with $\overline{W}_t$ based on model (2.6). The results shows that within sample mean is significant at $\alpha=0.05$.

Table 2. Logistic regression coefficient point estimates and standard erros and approximate 95% confidence interval for the final model to estimate probability giving two BMD measurements

Predictor	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	C.I. of $\hat{\beta}_i$
Intercept	-0.090	0.226	(-0.544, 0.364)
Other Region	-0.115	0.190	(-0.538, 0.227)
Black	0.533	0.190	(0.152, 0.915)
MAmer	0.200	0.268	(-0.338, 0.738)
Other	0.766	0.435	(-0.108, 0.641)
ORegion*Other	-1.444	0.560	(-2.569, 0.319)
Female	-0.109	0.116	(-0.342, 0.125)
BlackF	-0.662	0.208	(-1.081,-0.244)
MAmerF	-0.229	0.256	(-0.743, 0.285)
OtherF	0.669	0.777	(-0.892, 2.230)
Age20	0.355	0.225	(-0.096, 0.806)
Age30	0.277	0.213	(-0.151, 0.705)
Age40	0.929	0.246	(0.434, 1.424)
Age50	0.601	0.262	(0.074, 1.128)
Age60	0.383	0.200	(-0.019, 0.784)

Table 3. Summary of survey weights

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. Survey weighted regression coefficient estimates and standard errors and approximate 95% confidence intervals for measurement error variance regression model $S_t^2=\gamma_0+\gamma_1\overline{W}_t+\kappa_t$

Predictor	$\hat{\beta}_i\times10^4$	$se(\hat{\beta}_i)\times10^4$	C.I. of $\hat{\beta}_i\times10^4$
Intercept	0.012	1.869	(3.799, 3.823)
$\overline{W}_t$	4.163	2.047	(0.049, 8.277)

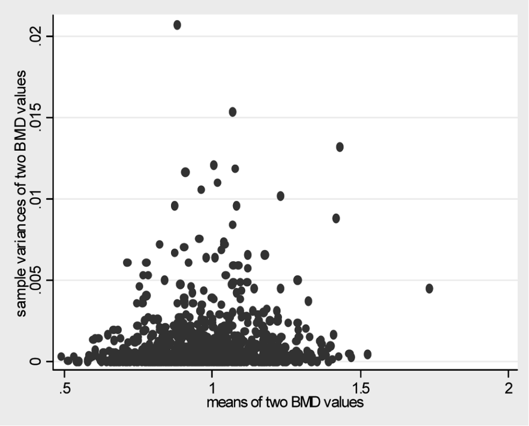

Fig. 1. Plot of sample mean $\overline{W}_t$ versus sample variances S_t^2 of two total BMD measurements

4. Conclusion

Measurement error is one of main concern to survey researchers with other sources of error. An observation measured with error can be expressed as a function of underlying true value plus error term. In some cases, the measurement error variance may also be a function of unknown true value, and the error variance function can be rewritten as a function of true value multiplied by a scale factor. This research has explored the estimation of linear measurement error variance function under complex sampling design when the scale factor is small. In the case, the linear measurement error variance can e estimated by the survey weighted regression estimators, and they are consistent estimators under some regularity conditions. The result was applied to the U.S. NHANES III data, which has replicate measurement for a small subset of original respondents's group. One measurement, total bone mineral density (BMD) measurement, was selected from the U.S. NHANES III data for empirical analysis, because the sample variances of total BMD measurement has a linearly increasing trend when the means of two measurements are increasing.

References

[1] M. P. Cooper, “Web Surveys: a Review of Issues and Approaches”. Public Opinion Quarterly, Vol. 64, pp. 464-494, 2000.

[2] T. E. Dalenius, “The survey statistician's reponsibility for both sampling and measurement errors”, in D. Krewski, R. Platek, and J. N. K. Rao (eds.), Current Topics in Survey Sampling, New York: Academic Press, pp. 17-29, 1981.

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

[4] M. Davidian and R. J. Carroll, “Variance function estimation”, J. Am. Stat. Assoc., Vol. 82, pp. 1097-1091, 1987.

[5] R. J. Carroll and D. B. H. Cline, “An asymptotic theory for weighted least -squares with weighted by replication”, Biometrics, Vol. 10, pp. 478-486, 1988.

[6] M. Davidian, “Estimation of variance functions in assays with possibly unequal replication and non-normal data”, Biometrika, Vol. 77, pp. 43-54, 1990.

[7] R. J. Carroll and L. A. Stefanski, “Approximate quasi-likelihood estimation in models with surrogate predictors”, J. Am. Statistical Assoc., Vol. 85, pp. 652-663, 1990.

[8] W. A. Fuller, “Regression estimation in the presence of measurement error”, in P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman 9eds.), Measurement Errors in Surveys, New York: John Wiley & Sons, pp. 616-635, 1991.

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

[10] J. L. Eltinge, S. Heo, and S. R. Lee, “Use of propensity models in the analysis of subsample re-measurements for NHANES III”, in Proceedings of the Survey Mehtods Section, Statistical Society of Canada, pp. 27-36, 1997.