Equivalence of the classical measurement errors model and the nested random effects model

Sangita Kulathinal*

July 17, 2007

Abstract

This paper compares the classical measurement error model to the nested random effects model and brings out the equivalence of the two models.

1 Introduction

The classical measurement errors model and the nested random effects model are part of the post-graduate courses. The underlying philosophy of the two models are quite different and hence, they are not usually covered in the same course. Nevertheless, their comparison is also not usually considered. During my teaching and consultancy, many researchers from the field of epidemiology asked me questions circling around these two models and unfortunately, I could not find a good reference discussing both these models. Hence, I decided to write it down myself.

2 Classical measurement error model

Experiments involving measurements are bound to have measurement errors. For example, systolic blood pressure, cholesterol, which are used as a predictor of coronary heart disease and exposure agents like dust, asbestos, temperature, rainfall, water-levels in wells. Many of these factors have strong daily and seasonal variations. In measuring these factors, various sources of error include simple machine recording error, administration error, time of the day and season of the year. It is well known that measurements with error, if used in the analysis, will result in misleading predictions of the response.

The classic model for measurement error is

$$W = X + U,$$

*Indic Society for Education and Development (INSEED), 1, Swami Enterprises Complex, Tigrania road, Nashik - 422 011, Maharashtra, India. e-mail: sangita.kulathinal@inseed.org

where W is the observed variable, X is the true variable and U is an additive error.

In practice, it is assumed that the error structure is known that is the probability distribution of U is known. Information about the error structure can be available from:

1. internal subset of the primary data, and/or
2. external or independent studies.

Three types of data can be available (may be only on the part of the original study group):

1. validation data, in which X is observed directly,
2. replication data, in which replicates of W are available,
3. instrumental data, in which another variable T is observable in addition to W.

One can employ any of the above-mentioned data set to characterise the error structure under following assumptions.

1. External data can be used provided the error structure in those data also applies to the present data. In many instances, approximately the same classical error model holds across different populations. So, parameters of a model can be transported from one study to another. In classical error model, it is reasonable to assume that the error distribution is the same across different populations (but not that of the true variable X).
2. One takes replicate measurements if there is a good reason to believe that the replicated mean is a better estimate of X than a single observation. In this set-up the true value denotes the long-term average of the variable. In this case, repeated measurements are used to estimate the variance of the measurement error, U.

The standard technique to take into account the measurement errors is regression calibration algorithm. The basic idea of this algorithm is to replace X by its estimate $E(X \mid W)$, regression of X on W using external or replicated data. Then use the standard techniques like regression models to study the effect of the factors on the response.

The estimation of the parameters of a linear regression mode1 where the variables were measured with error have been studied extensively over the last decade. We refer to references Fuller (1987) and Dear *et al.* (1997) and references therein. We now describe the model and give the estimate of true variable using replicated data. Let the data be described by the classical error model

$$W_{ij} = X_i + U_{ij}, j = 1, 2, ..., k_i, i = 1, 2, ..., n, \tag{1}$$

where k_i and X_i are the number of observations taken on individual i and the true long-term average of the variable for the individual i, $i = 1, 2, ..., n$. We

assume that $X_i \sim N(\mu, \sigma_x^2)$ and $U_{ij} \sim N(0, \sigma_u^2)$ for each i and j. Also, X and U are assumed to be independent of each other. The conditional expectation of X_i given the average of W_{ij} for fixed i is

$$E(X_i \mid \overline{W}_i) = \mu + \frac{\sigma_x^2}{\sigma_x^2 + \frac{\sigma_u^2}{k_i}}(\overline{W}_i - \mu), \tag{2}$$

where $\overline{W}_i = \sum_{j=1}^{k_i} W_{ij}/k_i$ and μ is the overall mean. The true value for the individual i can be estimated using the above conditional expectation but it involves unknown parameters μ, σ_x^2, and σ_u^2. The estimates of these unknown parameters are given below

$$\hat{\mu} = \frac{\sum_{i=1}^{n} k_i \overline{W}_i}{\sum_{i=1}^{n} k_i} \tag{3}$$

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_i} (W_{ij} - \overline{W}_i)^2}{\sum_{i=1}^{n} (k_i - 1)} \tag{4}$$

$$\hat{\sigma}_x^2 = \{[\sum_{i=1}^{n} k_i (\overline{W}_i - \hat{\mu})^2] - (n-1)\hat{\sigma}_u^2\}/v, \tag{5}$$

where $v = \sum_{i=1}^{n} k_i - (\sum_{i=1}^{n} k_i^2)/(\sum_{i=1}^{n} k_i)$. The above estimators are obtained using analysis of variance. Note that $\hat{\sigma}_u^2$ is actually the within-person variance and the first term in the expression of $\hat{\sigma}_x^2$ is the between-person variance. The true value of the variable is then estimated by replacing the unknown parameters by corresponding estimates.

3 Nested random effects model

The model generally used to study the changes in the factors over time using repeated measurements is nested random effects model (see Chambless *et al.*, 1992 and references therein). For the sake of completeness we will describe the nested random effects model here and bring out its comparison with the classical measurement error models in the next section. The model can be specified as

$$Y_{ijl} = \mu_i + \beta_{j(i)} + \epsilon_{l(ij)}, l = 1, 2, \ldots, m_j, j = 1, 2, \ldots, k_i, i = 1, 2, \ldots, n,$$

where the index l denote the number of measurements taken during visit j and in practice m_j is usually 2, k_i is the number of visits or equivalently number of observations on individual i, and that there are n such individuals. It is assumed that μ_i has $N(\mu, \sigma_{BP}^2)$ distribution. The term $\beta_{j(i)}$ denote the visit effect for the individual i. The term $\epsilon_{l(ij)}$ is the processing effect for the visit j for individual i. Both these effects are assumed to be normally distributed with mean zero. Also, the variables μ_i, $\beta_{j(i)}$, and $\epsilon_{l(ij)}$ are assumed to be independent of each other.

To model the variances of the two effects, either the standard constant variance or constant coefficient of variation assumptions are used. We will restrict our discussion to standard constant variance model where it is assumed that

$$variance(\beta_{j(i)}) = \sigma^2_{WP}$$

for fixed i and

$$variance(\epsilon_{l(ij)}) = \sigma^2_{\epsilon}$$

for fixed i and j. Here σ^2_{WP} denote the within-person variance and σ^2_{ϵ} denote the error variance due to method/model.

4 Comparison of the two models

Taking the average with respect to l in the nested random effects model, and writing $\overline{y}_{ij} = W_{ij}$, $\mu_i = X_i$, and $\beta_{j(i)} + \overline{\epsilon}_{(ij)} = U_{ij}$ and also $\sigma^2_{BP} = \sigma^2_x$, and $\sigma^2_{WP} + \sigma^2_{\epsilon}/2 = \sigma^2_u$, we have the classic measurement error model and the conditional expectation given above can be used to estimate the true value of the variable.

It is important to note that the measurement error variance is split into two parts and each component is estimated because of multiple laboratory measurements during each visit and also multiple visits of each individual.

To summarize, we express all the variances in terms of the between-person, within-person and error variances as

$$X_i = \hat{\mu} + \frac{\hat{\sigma}^2_x}{\hat{\sigma}^2_x + \frac{\hat{\sigma}^2_u}{k_i}}(\overline{W}_i - \hat{\mu}),$$

where $\hat{\mu}$ is an overall mean of the observations W, $\hat{\sigma}^2_x$ is the between-person variation/inter-individual variance, $\hat{\sigma}^2_u$ is the sum of the within-person variation/intra-individual variation and half of the error variation (which includes method and errors). Note that the error variance is divided by 2 only if we have two observations at each visit and the average of the two is used. For those individual who has only single visit, we estimate X by

$$X_i = \hat{\mu} + \frac{\hat{\sigma}^2_x}{\hat{\sigma}^2_x + \hat{\sigma}^2_u}(W_i - \hat{\mu}).$$

If only one measurement is taken during each visit then $\hat{\sigma}^2_u$ is the sum of the within-person variation/intra-individual variation and the error variation. Note that the two variations can not be then estimated separately.

5 Concluding remarks

In this paper, we presented the equivalence of the classical measurement errors and nested random effects model. Both these models are used in many disciplines. It is well-known that if the data with measurement errors are used

without correcting for the errors then the effects are diluted. Using the values of various variance components, one can obtain the estimates of the true factors using the equation given in the earlier section and use them in appropriate regression model to adjust for the regression dilution, which might occur due to measurement error in these factors. None of the papers which claim to have been correcting for regression dilution describe the procedure for estimating the true values of the factors. This is an attempt to put two models together and suggest a way to estimate the true variables using repeated measurements.

The Cox's proportional hazards model is commonly used in analysing data from epidemiologic studies. It is shown by Prentice (1982), Pepe *et al.* (1989) and others that estimation of regression parameters in Cox's model reduces to regression calibration problem in the situations when the event is rare and the conditional distribution of X given W is normal. Clayton (1991) suggested regression calibration method for each risk set and hence relaxing the assumption of rare event.

It will be interesting to see how one can estimate the regression coefficients in Cox's proportional hazard model directly when the explanatory variables are measured with errors. This approach will be different compared to the usual approach of estimating the true values using regression calibration and then carry out the standard technique of Cox's model using these estimates.

References

Chambless L.E., McMahon R., Wu K., Folsom A., Finch A., Shen Y.L., 1992, Short-term intraindividual variability in hemostasis factors. The ARIC study, *Ann Epidemiol*, **2**, 723-733.

Clayton, D.G., (1991), Models for the analysis of cohort and case-control studies with inaccurately measured exposures, in Statistical Models for Longitudinal Studies of Health (Monographs in Epidemiology and Biostatistics, Vol 16), J.H. Dwyer, M. Feinleib, P. Lippert, H. Hoffmeister, Oxford University Press, New York, pp. 301-331.

Dear K.B.G., Putennan M.L., Dobson A.J., 1997, Estimating corre1ations from epidemio1ogica1 data in the presence of measurement error, *Statistics in Medicine*, **16**, 2177-2189.

Fuller W.A., 1987, *Measurement Error Mode/s* (Wi1ey: New York).

Pepe, M.S., Self, S.G. and Prentice, R.L., 1989, Further results in covariate measurement errors in cohort studies with time to response data, *Stat. in Med.*, **8**, 1167-1178.

Prentice, R.L., 1982, Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika*, **69**, 331-342.