Multilevel Linear Regression Analysis of Complex Survey Data

Fan Zhang, Sameena Salvucci, Synectice for Management Dec. Inc, Arlington, VA
Mike Cohen, National Center for Education Statistics, Washington DC

Key Words: Multilevel Linear Regression, Design Based Regression Analysis, Hierarchical Linear Model, Selection Effect, HLM.

1. Introduction

Multilevel (hierarchical) linear regression analysis is often applied in social research, especially educational research due to an increasing awareness that many problems in educational research have multilevel characteristics and due the numerous recent related studies and the development of computer software. The theoretical and the computing development, however, although very appealing, is mostly limited within the classical framework, i.e., assuming the data are generated by the underlying multilevel linear model and no complex sample selection effects are taken into account.

As the multilevel (hierarchical) linear modeling technique gets widely applied, researchers realized that the data available for the application are rarely that simple. The data almost universally come from large-scale complex surveys. And the sample selection procedure of the survey almost always has features such as multistage, unequal probability selection, clustering, etc. The realized sample is the product of both underlying multilevel model and sample selection procedure. Ignoring the selection effects may cause bias in both point estimate and variance estimates.

Different approaches have been taken to tackle the problem. One approach is to use the standard multilevel regression software (HLM, for example) and apply sampling weights since the software provides this option. In this approach, researchers can take the multilevel error structure the data present into account and still correct the point estimate bias caused by the unequal probability selection. In another approach, the researchers realize that the effect of sampling selection may be more than just biased point estimates. The variance estimates can be also biased too. But since the standard multilevel regression software does not specifically calculate the variance caused by the sample selection, the researchers turn to the available one level design based regression software such as SUDAAN and WESVAR. In this approach, the level-2 data (school data, for example) are merged to level-1 data (student data, for example) and a one level design based regression analysis is performed. Here the sampling selection effect is well addressed but the multilevel error structure is given up. Therefore it might turn out to be less efficient than this is also incorporated into the estimation.

In this study, we discuss the efficacy of these different approaches in terms of the biases of point estimates and related variance estimates. In section 2 we introduce multilevel linear regression models. We can then see that the model coefficients γ (including the school effects) can be also estimated by one level model. In section 3, we discuss the efficacy of multilevel modeling vs. one level modeling within the classical framework, assuming no selection effects. In section 4, we examine the efficacy of the one level design based regression approach. In section 5, we examine the weighted multilevel regression approach and discuss the conditions that this approach stands.

2. Hierarchical (Multilevel) Linear Model

A good example of a hierarchical structure is an educational system where students are "clustered" or grouped within classes. To reflect this in the model, assume there are m groups, indexed by j, and there are n_j of individuals in group j. Let

$$\underline{y}_j = (\underline{y}_{1j}, \underline{y}_{2j}, \cdots \underline{y}_{nj})',$$

$$\underline{\beta}_j = (\underline{\beta}_{1j}, \underline{\beta}_{2j}, \cdots \underline{\beta}_{nj})',$$

$$\mathbf{X_j} = \begin{bmatrix} 1 & X_{1j1} & \cdots & X_{1j,P-1} \\ 1 & X_{2j1} & \cdots & X_{2j,P-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n_j j1} & \cdots & X_{n_j j,P-1} \end{bmatrix}.$$

The micro level (or level 1) model is

$$(2.1) \qquad \underline{\mathbf{y}}_j = \mathbf{X_j}\underline{\beta}_j + \underline{\varepsilon}_j .$$

Also let

$$\mathbf{z'_j} = (1, Z_{j1}, \cdots Z_{j,Q-1}),$$

$$\gamma_p = (\gamma_{op}, \gamma_{1p}, \cdots \gamma_{Q-1,p})'$$

$$\mathbf{Z_j} = \begin{bmatrix} \mathbf{z}'_j & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}'_j & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{z'_j} \end{bmatrix}_{P\times PQ} , \gamma = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{P-1} \end{pmatrix}_{PQ\times 1} .$$

The macro level (or level 2) model is

(2.2) $\underline{\beta}_j = \mathbf{Z}_{\mathrm{j}}\gamma + \underline{\delta}_j$.

It is assumed that all entries of $\underline{\varepsilon}_j$ are independent of each other and different $\underline{\varepsilon}_j$ are independent of each other and different $\underline{\delta}_j$ are independent of each other. It is also assumed that the macro level disturbances $\underline{\delta}_j$ are independent of the micro level disturbances $\underline{\varepsilon}_j$. For the expectation and covariance matrix, it is often assumed that $E(\underline{\varepsilon}_j) = \mathbf{0}$, $V(\underline{\varepsilon}_j) = \sigma^2\mathbf{I}$, $E(\underline{\delta}_j) = \mathbf{0}$ and

$$V(\underline{\delta}_j) = \Omega = \begin{bmatrix} \tau_{00} & \tau_{01} & \cdots & \tau_{0,P-1} \\ \tau_{10} & \tau_{11} & \cdots & \tau_{1,P-1} \\ \vdots & \vdots & \cdots & \vdots \\ \tau_{P-1,0} & \tau_{P-1,1} & \cdots & \tau_{P-1,P-1} \end{bmatrix}.$$

Model (2.1) and (2.2) can be combined into:

(2.3) $\underline{\mathbf{y}}_j = \mathbf{X}_{\mathrm{j}}\mathbf{Z}_{\mathrm{j}}\gamma + \mathbf{X}_{\mathrm{j}}\underline{\delta}_j + \underline{\varepsilon}_j$.

A more compact form can be obtained as following: let $\mathbf{X} = \mathbf{X}_1 \dotplus \ldots \dotplus \mathbf{X}_m$, (the direct sum of $\mathbf{X}_{\mathrm{j}}$'s, $\mathbf{X}$ is an $n \times mP$ block diagonal matrix with $\mathbf{X}_1, \ldots, \mathbf{X}_m$ as the diagonal blocks), stack $\underline{\mathbf{y}}_j$ on top of each other to form the n-vector $\underline{\mathbf{y}}$, stack $\underline{\varepsilon}_j$ and $\underline{\beta}_j$ in the same way to form $\underline{\varepsilon}$ and $\underline{\beta}$, that is

$$\underline{\mathbf{y}} = \begin{pmatrix} \underline{\mathbf{y}}_1 \\ \underline{\mathbf{y}}_2 \\ \vdots \\ \underline{\mathbf{y}}_m \end{pmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_m \end{bmatrix}, \underline{\beta} = \begin{pmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \\ \vdots \\ \underline{\beta}_m \end{pmatrix},$$

$\underline{\varepsilon} = (\underline{\varepsilon}_1, \underline{\varepsilon}_2, \cdots \underline{\varepsilon}_m)'$. The combined matrix form is:

(2.4) $\underline{\mathbf{y}} = \mathbf{XZ}\gamma + \mathbf{X}\underline{\delta} + \underline{\varepsilon}$,

with variance-covariance matrix

(2.5) $Var(\underline{\mathbf{y}}) = \mathbf{X}\Sigma\mathbf{X}' + \sigma^2\mathbf{I} = \mathbf{V}\sigma^2$.

Here $\Sigma = \begin{bmatrix} \Omega & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Omega & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Omega \end{bmatrix}$. Assume V is nonsingular.

A very important consequence of the model (2.4) is the loss of independence among the entries of $\underline{y}_j$. This can be seen from the variance-covariance matrix of $\underline{y}_j$: $Var(\underline{y}_j) = Var(\mathbf{X}_{\mathrm{j}}\underline{\delta}_j + \underline{\varepsilon}_j) = \mathbf{X}_j\Omega\mathbf{X}'_j + \sigma^2\mathbf{I}$. It is easy to see the off diagonal entices of the variance-covariance matrix usually are not zeros.

When there are no random effects in the macro level model (2.2), the hierarchical linear model reduces to the ordinary regression model that includes micro-level variables, X_{ij}, macro-level variables, Z_{qj} and their interaction terms. Actually the model becomes

(2.5) $\underline{\mathbf{y}} = \mathbf{XZ}\gamma + \underline{\varepsilon}$.

3. One Level Model vs. Multilevel Level Model

One level linear regression model (2.5) is often used in school effects research especially before 1980's. Usually the school characteristics are merged to student data (students within the same school are merged with the same school level variables), then treat the school level variables and the student level variables as they are from the same level and perform the ordinary one level regression analysis. If the data structure exhibits intra-cluster correlation and the analyst fails to take account of the correlation structure in the statistical model, the underlined model is misspecified. In this section, we study the effect of this model misspecification.

First consider the combined matrix form (2.4) that correctly models the data. Although now the observations are not independent of each other, it can be shown that the minimum variance unbiased estimator (BLUE) of γ is

(3.1) $\hat{\gamma}_{GLS} = (\mathbf{Z}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{XZ})^{-1}\mathbf{Z}'\mathbf{X}'\mathbf{V}^{-1}\underline{\mathbf{y}}$.

It is easy to see $\hat{\gamma}_{GLS}$ is unbiased. That is

$$E(\hat{\gamma}_{GLS}) = (\mathbf{Z}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{XZ})^{-1}\mathbf{Z}'\mathbf{X}'\mathbf{V}^{-1}E(\underline{\mathbf{y}}) = \gamma.$$

And the variance - covariance matrix of $\hat{\gamma}_{GLS}$ is

(3.2) $Var(\hat{\gamma}_{GLS}) = (\mathbf{Z}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{XZ})^{-1}\sigma^2$.

In reality, $\mathbf{V} = (1/\sigma^2)\mathbf{X}\Sigma\mathbf{X}' + \mathbf{I}$ is never known. Various methods and software (HLM and MP3, for example) were developed to estimate the parameters.

In another hand, if the correlation structure is overlooked and the model is misspecified, that is, if the second level random effects $\underline{\delta}$ are set to zero and OLS is applied to the ordinary regression model $\underline{\mathbf{y}} = \mathbf{XZ}\gamma + \underline{\varepsilon}$, the resulting estimator of γ is the ordinary least square (OLS) estimator

(3.3) $\hat{\gamma}_{OLS} = (\mathbf{Z}'\mathbf{X}'\mathbf{XZ})^{-1}\mathbf{Z}'\mathbf{X}'\underline{\mathbf{y}}$.

$\hat{\gamma}_{OLS}$ is also unbiased for γ. That is

$$E(\hat{\gamma}_{OLS}) = (\mathbf{Z'X'XZ})^{-1}\mathbf{Z'X'}E(\underset{\sim}{\mathbf{y}}) = \gamma .$$

But the associated variance-covariance matrix becomes

(3.4) $Var(\hat{\gamma}_{OLS}) = (\mathbf{Z'X'XZ})^{-1}\mathbf{Z'X'VXZ}(\mathbf{Z'X'XZ})^{-1}\sigma^2$.

Compare (3.4) with (3.2), we see in general elements of (3.4) would provide larger variances both for individual coefficients and for linear functions of the coefficients. Therefore, both $\hat{\gamma}_{GLS}$ and $\hat{\gamma}_{OLS}$ are unbiased for γ, but $\hat{\gamma}_{OLS}$ is less efficient. For a special case of two level model, Scott & Holt (1982) showed that

$$1 \le \frac{Var(\mathbf{c}'\hat{\gamma}_{OLS})}{Var(\mathbf{c}'\hat{\gamma}_{GLS})} \le \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n},$$

here $\mathbf{c}'\hat{\gamma}_{OLS}$ is any linear combination of the regression coefficients $\hat{\gamma}_{OLS}$, $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ are the eigenvalues of $\mathbf{V}$. Their conclusion is that the loss of efficiency is not usually worth worrying about when compared with the other usual problems that we encounter with survey data, such as nonresponse, adequacy of the model, and so on.

The application of OLS estimator may cause another problem. In standard software programs such as SAS and SPSS, when OLS is applied, the estimate of $Var(\hat{\gamma}_{OLS})$ can be seriously misleading. The estimate of $Var(\hat{\gamma}_{OLS})$ produced from these programs is

(3.5) $\hat{V}(\hat{\gamma}_{OLS}) = (\mathbf{Z'X'XZ})^{-1}\hat{\sigma}^2_{OLS}$,

with $\hat{\sigma}^2_{OLS} = (\underset{\sim}{\mathbf{y}} - \mathbf{XZ}\hat{\gamma}_{OLS})'(\underset{\sim}{\mathbf{y}} - \mathbf{XZ}\hat{\gamma}_{OLS})/(n-P)$.

Compare (3.5) to (3.4) we can see that $\hat{V}(\hat{\gamma}_{OLS})$ can be a badly biased estimate of the true variance matrix. As in the special case shown by Scott and Holt (1982) where $E(\hat{\sigma}_{OLS}) \cong \sigma^2$, the difference between (3.4) and $E[\hat{V}(\hat{\gamma}_{OLS})]$ is factor $(\mathbf{Z'X'XZ})^{-1}\mathbf{Z'X'VXZ}$.

4. Design-Based One Level Regression Analysis of Multilevel Survey Data

In the previous sections, we assumed that the sample is generated by a hypothetical sequence of replications of an experiment described by multilevel model (2.4). And our goal is to estimate and inference γ. In reality, however, the sample is often selected from a larger existing finite population of the data in stead of a sequence of replications of an experiment. Usually the selection of the observed sample is performed according to a probability sampling design. Efficiency and administration consideration and complex population structure often lead to a complex sampling design for the sample selection procedure. For example, the finite population may be stratified according to some auxiliary variable known to all the units in the population and units in different strata are selected with different probability.

Therefore, the observed sample selected according to a complex sampling design can be viewed as a product of a two step procedure (Hartley and Sielken, 1975):

1. An N unit population is generated by a hypothetical sequence of N replications of an experiment described by a model such as (2.4);
2. An $n < N$ unit sample is selected from the population of size N obtained in step 1 according to a complex survey design.

The complex sample selection in step 2 adds another stochastic mechanism to the realized sample. So now there are two random components to be taken into account: one is the distribution described by model (2.4), also referred as ξ distribution, and the second one is the randomization distribution, also referred as D distribution.

Failed to take the randomization distribution into account, the two estimators of γ considered in the previous section, $\hat{\gamma}_{GLS}$ and $\hat{\gamma}_{OLS}$ both are biased in general when selection effect presents (see for example, Nathan and Holt,1980).

In practice, in order to take the randomization into account and also due to the limitation of available computing software, analysts often merge the level 2 data to the level 1 data. And conduct a one level design based regression analysis using software such as SUDAAN that can incorporate the effect of complex sampling. The design based one level regression approach uses the following weighted ordinary least square estimator to estimate γ:

(4.1) $\hat{\gamma}_{WOLS} = (Z_s' X_s' \Pi_s^{-1} X_s Z_s)^{-1} Z_s' X_s' \Pi_s^{-1} y_s$,

where $\Pi_s = diag(\pi_1, \ldots, \pi_n)$ and π_i is the probability that unit *i* is selected into the sample *s*. And the design variance of $\hat{\gamma}_{WOLS}$ is estimated. In other words, the variance with respect to the randomization is estimated and used as the variance estimate of $\hat{\gamma}_{WOLS}$. This approach can be justified as following. First, $\hat{\gamma}_{WOLS}$ is approximately unbiased with respect to both model and randomization distribution. Actually, notice

$E_D(\hat{\gamma}_{WOLS}) \approx (Z_N' X_N' X_N Z_N)^{-1} Z_N' X_N' y_N = \hat{\gamma}_{WOLS,N}$,

here subscript N represent the N units in the finite population (see, for example, Särndal *et al* 1991). Since the *N* units in the finite population are outcomes of *N* replications of the model, $\hat{\gamma}_{WOLS,N}$ is unbiased under ξ distribution for γ. Therefore $E_\xi E_D(\hat{\gamma}_{WOLS}) \approx \gamma$.

For the variance estimation, notice we may decompose the variance of $\hat{\gamma}_{WOLS}$ as (see for example, Pfeffermann 1993)

$$(4.2) \qquad Var_{D\xi}(\hat{\gamma}_{WOLS}) = E_\xi[Var_D(\hat{\gamma}_{WOLS} \mid y_N)] + \\ + Var_\xi[E_D(\hat{\gamma}_{WOLS} \mid y_N)] \\ = E_\xi[Var_D(\hat{\gamma}_{WOLS} \mid y_N)] + O(N^{-1}).$$

Therefore when the population size N is large the variance of $\hat{\gamma}_{WOLS}$ under the $D\xi$ distribution can be estimated by randomization variance $Var_D(\hat{\gamma}_{WOLS} \mid y_N)$. Methods of estimating $Var_D(\hat{\gamma}_{WOLS} \mid y_N)$ have been developed (see, for example, Särndal *et al* 1991) and software is available (e.g. SUDAAN, WesVar PC).

5. Multilevel Linear Regression Analysis of Multilevel Survey Data

In the one level design based approach described in the above section, the multilevel error structure of the data is not taken into account in the modeling and estimation. According to the discussion in Section 3, incorporating the multilevel error structure into the estimation may result in more efficient estimator. To this end, different approaches are taken in practice. The first approach we shall discuss here is ordinary multilevel regression analysis (section 5.1), which completely ignores selection effect. The second approach we shall discussed is the weighted multilevel regression analysis (section 5.2), which incorporates the unequal selection probability but does not calculate the randomness variance. The third approach considered (section 5.3) not only use design based approximately unbiased point estimator but also estimates its randomness variance.

5.1 Ordinary Multilevel Regression Analysis of Multilevel Survey Data

When ordinary multilevel regression analysis is performed to complex survey data, the multilevel survey data are treated as if they are n replicates of model (4.1) and the multilevel error structure is taken into account in the estimation but the sampling design is totally ignored. Here $\hat{\gamma}_{GLS} = (\mathbf{Z}'_s\mathbf{X}'_s\mathbf{V}_s^{-1}\mathbf{X}_s\mathbf{Z}_s)^{-1}\mathbf{Z}'_s\mathbf{X}'_s\mathbf{V}_s^{-1}y_s$ is used to estimate γ and $\hat{V}_\xi(\hat{\gamma}_{GLS})$, the variance estimate of $\hat{\gamma}_{GLS}$ with respect to the model, is used to estimate the total variance. For this approach, both point estimator and variance-covariance estimators can be biased. For the point estimator, as Nathan and Holt 1980 have pointed out, estimator $\hat{\gamma}_{GLS}$ is in general biased for γ under $D\xi$ distribution. One argument for this approach and the approach described in the next subsection 5.2 is that the multilevel nature of hierarchical linear model directly models the multilevel sampling design used in the data collection. Therefore the modeling that is performed in the commonly used hierarchical linear model software accurately reflects the sampling design. This is a misunderstanding of the relationship between population structure and design. In one hand it is the population structure, as described by the model such as (2.4), that generates the multilevel error data. Therefore, the analyst must take account of the population structure in the statistical model no matter what design has been is used. In another hand, however, different designs generate different sample spaces that in turn will affect the statistical properties of the underlying estimators. As for the variance-covariance estimator, even in the case where the specified population structure coincides with the multilevel sampling design, e.g. sampling students within sampled schools, the estimating performed in the standard hierarchical linear model software such as HLM or HLMPV does not necessarily reflect the randomness variance. Failed to estimate the randomness variance may lead to bias in the variance-covariance matrixes of $\hat{\gamma}_{GLS}$ and $\hat{\gamma}_{WGLS}$, the weighted version of $\hat{\gamma}_{GLS}$ which we will discuss in the next subsection.

Next, we consider the condition that sampling design can be ignored in the reference of γ when ordinary multilevel regression analysis is performed. And we only consider the maximum likelihood estimator since the generalized least square estimator, maximum likelihood and empirical Byes estimator are actually all identical (Raudenbush, 1984).

When maximum likelihood method is used to estimate γ and the sampling design is completely ignored, the likelihood is based only on the conditional distribution of the sampled y's with the sample s fixed and ignore all other random components. That is the likelihood is based on the following density function of the observed data in order to obtain the maximum likelihood estimator of γ:

$$k_s(y_s, x_s, z_s; \gamma_1, \alpha) = f_s(y_s \mid x_s, z_s; \gamma_1) h_s(x_s, z_s; \alpha),$$

or equivalently based on conditional density function

$$(5.1) \qquad f_s(y_s \mid x_s, z_s; \gamma_1).$$

The likelihood function based on (5.1) is sometimes called the face-value likelihood since it is not based on the full distribution of all random components. As we saw in section 4, the realized sample can be viewed as a product of a two-step procedure. One is the replication of the model and the other one is the probability sampling. In probability sampling, the sample selection only depends on certain design variables that are differ from the response variable y. In other words in probability sampling the designs can be written as $p(s \mid u), s \in \aleph$. Here s is the realized sample, $\aleph$ is the set of all feasible samples, and u are the design variables, which may include label information such as cluster or stratum indicator variables which determine group membership, other group variables and quantitative variables such as measures of size. In reality, however, the design information known to the analyst is usually only partial design information, which can be denoted as a function of u: $d_s = D_s(u)$. Here $d_s = D_s(u)$ is data derived from knowledge of the selection mechanism and from values of the selection probabilities and also from any known values or functions of u (Sugden and Smith, 1984). Therefore, the full likelihood should be based on the joint density function of (s, y_s, x_s, z_s, d_s):

$$g(s, y_s, x_s, z_s, d_s; \phi, \gamma_2, \varphi)$$
$$= \int_{D_s} p(s \mid y_s, x_s, z_s, u; \phi)\, f_s(y_s \mid x_s, z_s, u; \gamma_2)\, h_s(x_s, z_s, u; \varphi)\, du\,.$$

Here y_s are the observed y's and $D_s = \{u : D_s(u) = d_s\}$. Since for all probability samplings the designs only depend on the design variable u, that is $p(s \mid w, u; \phi) = p(s \mid u)$ for any variable w and parameter ϕ, we can rewrite the above joint density function as

$$(5.2)\quad g(s, y_s, x_s, z_s, d_s; \phi, \gamma_2, \varphi)$$
$$= \int_{D_s} p(s \mid u)\, f_s(y_s \mid x_s, z_s, u; \gamma_2)\, h_s(x_s, z_s, u; \varphi)\, du\,.$$

Compare (5.1) and (5.2) we can see that a sufficient condition of complete ignorance of design information is that y_s and u are independent conditional on (x_s, z_s). That is

$$(5.3)\qquad f_s(y_s \mid x_s, z_s, u; \gamma_2) = f_s(y_s \mid x_s, z_s; \gamma_1)\,.$$

An interesting special case where (5.3) stands is when all design variables u are included in the model predictor x and/or z. Let x_u and z_u be the design variables that are also level-1 or level-2 predictors respectively, that is $u = (x_u, z_u)$. Notice since we assume all design variables are included in the model, the columns of x_u and z_u are subsets of the columns of x_s and z_s. Although u has rows for all N units in the finite population and x_s and z_s only have rows for the n units in the sample s, this information is redundant for y_s given (x_s, z_s). Therefore,

$$f_s(y_s \mid x_s, z_s, u; \gamma_2) = f_s(y_s \mid x_s, z_s, x_u, z_u; \gamma_2)$$
$$= f_s(y_s \mid x_s, z_s; \gamma_1)\,.$$

Situation where design variables coincide with predictors is not rare in educational surveys. For example, school sector (public/private) is often used as a first stage stratification variable and student ethnicity is often used as a second stage stratification variable and both variables are often used as predictors in multilevel regression analysis. However, it should be pointed out that in reality usually only partial design variables are included in the model so the design can still be informative for the inference of γ.

5.2 Weighted Ordinary Multilevel Regression Analysis of Multilevel Survey Data

This approach is often seen when analysts use standard multilevel regression software such as HLM or HLM-pv and apply sampling weights in the analysis. In this approach, unequal selection probability and the multilevel error structure are taken into account in the parameter estimation but the randomness variance is not estimated. Here the finite population parameter to be estimated is $(Z'_N X'_N V_N^{-1} X_N Z_N)^{-1} Z'_N X'_N V_N^{-1} y_N$. Obviously this quantity is unbiased for γ under ξ distribution. It is not clear to us what exact estimator is used to estimate this quantity in HLM. But Pfeffermann & LaVange 1989 propose the following estimator:

$$\hat{\gamma}_{WGLS} = \sum_{c=1}^{m} \frac{1}{\pi_c} \left[X_c^{*\prime} W_c X_c^{*} - X_c^{*\prime} W_c X_c Q_{c,W}^{-1} X'_c W_c X_c^{*} \right]^{-1} R$$

where $Q_c = X'_c X_c + \sigma^2 \Delta^{-1}$. π_c is the cluster c inclusion probability. And $W_c = diag(w_{c1}, \cdots w_{cn_c})$ is the matrix of sampling weights corresponding to sampling within the cth selected cluster. $Q_{c,w} = X'_c W_c X_c + \sigma^2 \Delta^{-1}$, $X_c^{*} = X_c Z_c$ and

$$R = \sum_{c=1}^{m} \frac{1}{\pi_c} \left[X_c^{*\prime} W_c Y_c - X_c^{*\prime} W_c X_c Q_{c,W}^{-1} X'_c W_c Y_c \right].$$

Pfeffermann and LaVange 1989 claim $\hat{\gamma}_{WGLS}$ is approximately unbiased under D distribution for $(Z'_N X'_N V_N^{-1} X_N Z_N)^{-1} Z'_N X'_N V_N^{-1} y_N$. Therefore it is also approximately unbiased for γ under $D\xi$ distribution. As for the variance estimation, similarly the variance of $\hat{\gamma}_{WGLS}$ can be written as

$$Var_{D\xi}(\hat{\gamma}_{WGLS}) = E_{\xi}[Var_D(\hat{\gamma}_{WGLS} \mid y_N)]$$
$$+ Var_{\xi}[E_D(\hat{\gamma}_{WGLS} \mid y_N)].$$

In the standard ordinary multilevel regression software, the estimated variance of $\hat{\gamma}_{WGLS}$ is not the design based variance component $Var_D(\hat{\gamma}_{WGLS} \mid y_N)$. Hence the estimated variance-covariance matrix can be biased.

In this approach, partial design information (design weight) is taken into account in the estimation. In the following, we consider the conditions under which partial design information is sufficient to ignore the selection effect for the reference of γ.

When partial design information is used in the maximum likelihood estimation but the selection is ignored, the likelihood function is based on the following density function

$$k_s(y_s, x_s, z_s, d_s; \gamma_1, \alpha,).$$
$$= \int_{D_s} f_s(y_s \mid x_s, z_s, u; \gamma_1)\, h_s(x_s, z_s, u; \alpha)\, du .$$

However, the full distribution of the data (s, y_s, x_s, z_s, d_s) is still (5.2). Apply a result of Sugden & Smith 1984, the condition for the ignorability of sampling selection can be written as

$$f_s(y_s \mid x_s, z_s, u; \gamma_1) = f_s(y_s \mid x_s, z_s, d_s; \gamma_1).$$

In other words, observed y's are independent of full design variable vector u given the partial design information d_s. When this condition stands, we have

$$k_s(y_s, x_s, z_s, d_s; \gamma_1, \alpha,)$$
$$= f_s(y_s \mid x_s, z_s, d_s; \gamma_1) \int_{D_s} h_s(x_s, z_s, u; \alpha)\, du$$

and

$$g(s, y_s, x_s, z_s, d_s; \gamma_1, \varphi)$$
$$= f_s(y_s \mid x_s, z_s, d_s; \gamma_1) \int_{D_s} p(s \mid u)\ h_s(x_s, z_s, u; \varphi)\, du$$

Therefore, the likelihood functions based on both density functions produce the same maximum likelihood estimator of γ_1. Sugden and Smith 1984 examine some standard probability designs to see to what extent they satisfy the conditions for ignorability. It should be pointed out that in general knowing the inclusion probabilities even for all units is not sufficient to ignore the sampling design.

5.3 Design Based Multilevel Regression Analysis of Multilevel Survey Data

This approach is an improvement of the approach described in subsection 5.2. Here $\hat{\gamma}_{WGLS}$ is used as estimator of γ to take the multilevel error structure into account and to obtain an approximately unbiased estimate of γ under the joint $D\xi$ distribution. And the randomness variance is estimated and incorporated into the variance estimates. An example of this approach can be found in Pfeffermann et al 1998, in which design variance is estimated by Taylor Linearization method and used as estimate of the total variance.

Although with this approach the multilevel error structure and the selection effect both are taken cared of very well, $\hat{\gamma}_{WGLS}$ is not necessary more efficient than $\hat{\gamma}_{WOLS}$. Compare with that in section 3 we showed $\hat{\gamma}_{GLS}$ is more efficient than $\hat{\gamma}_{OLS}$ when there is no selection effect. In other words, this approach is not necessarily more efficient than the design based one level regression approach. Actually, notice

$$Var_{D\xi}(\hat{\gamma}_{WGLS}) = E_\xi[Var_D(\hat{\gamma}_{WGLS} \mid y_N)] + O(N^{-1})$$

and

$$Var_{D\xi}(\hat{\gamma}_{WOLS}) = E_\xi[Var_D(\hat{\gamma}_{WOLS} \mid y_N)] + O(N^{-1}).$$

Here both $Var_D(\hat{\gamma}_{WGLS} \mid y_N)$ and $Var_D(\hat{\gamma}_{WOLS} \mid y_N)$ depend on sampling design. Therefore, efficiency must be considered with respect to the specific sampling design. And this needs to be further investigated.

Reference:

Nathan, G. and Holt, D (1980): The effect of survey design on regression analysis. *J. R. Statist. Soc. B,* 42, No.3, pp. 377-386.

Pfeffermann, D. and LaVange, L. (1989): Regression models for stratified multi-stage cluster samples. *Analysis of Complex Surveys*, edited by Skinner, C.J., Holt, D., and Smith, T.M.F. pp. 237-260.

Pfeffermann, D. (1993): The role of sampling weights when modeling survey data. *International Statistical Review* , 61, 2, pp. 317-337.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., Rasbash, J. (1998): Weighting for unequal selection probabilities in multilevel models. *J. R. Statist. Soc. B*, 60, part 1, pp23-40.

Raudenbush, S. (1984): Applications of a hierarchical linear model in educational research. Ph.D. Dissertation.

Särndal , C-E, Swensson, B., Wretman, J. (1991): *Model Assisted Survey Sampling*. Springer-Verlag New York, Berlin, Heidelberg.

Sugden, R. A., Smith, T. M. F. (1984): Ignorable and informative designs in survey sampling inference. Biometrika, 1984, 71, 3, pp. 495-506.