Constructing Covariates in Multilevel Regression

Tihomir Asparouhov
Muthen & Muthen

Bengt Muthen
UCLA

Mplus Web Notes: No. 11

February 15, 2006
Version 2, March 19, 2007

[1]The authors are thankful to Oliver Ludtke, Herbert W. Marsh, Alexander Robitzsch and Ulrich Trautwein for finding and correcting an error in formula (10) in the previous version of this note.

1 Overview

Suppose that Y_{ij} is an observed dependent variable for individual i in cluster j and suppose that X_{ij} is an observed predictor for that individual. The standard two level regression model is concerned with determining the effect the variable X has on the dependent variable Y, however in multilevel settings we are also interested in distinguishing between purely individual effects and a group effect, see Raudenbush & Bryk (2002) page 139-141. For example if Y_{ij} is a student performance measurement and X_{ij} is the student socioeconomic status (SES) for student i in school j we would be interested in the effect of the school average SES and also the effect of the individual SES (individual deviation from the average). It is conceivable that the school average SES effect could be positive while the individual SES effect could be zero or even negative. Often a small number of students are sampled from each school and thus the individual SES variable is available for a small number of students. The school average SES variable we would like to use in the model is not available but it is often approximated by the average of the sampled students SES variables

$$\bar{X}_{.j} = \frac{1}{l}\sum_{i=1}^{l} X_{ij}$$

where l is the number of sampled units from cluster j. The two level regression model with separate individual and group level effects is then given by (see the group-mean centered model of Raudenbush-Bryk (2002), Table 5.11, page 140)

$$Y_{ij} = \beta_{0j} + \beta_w(X_{ij} - \bar{X}_{.j}) + \varepsilon_{ij} \quad (1)$$

$$\beta_{0j} = \mu + \beta_b\bar{X}_{.j} + \varepsilon_j \quad (2)$$

In this note we discuss the implications of approximating the true cluster average covariate by the sample cluster average covariate $\bar{X}_{.j}$. We describe three alternative estimation approaches available in Mplus for models that include cluster average covariates.

Each variable in two-level settings can be decomposed as the sum of the cluster average plus the individual deviation from the cluster average. For example for the dependent variable Y we have

$$Y_{ij} = Y_{jb} + Y_{ijw} \quad (3)$$

where Y_{jb} is the cluster average variable and Y_{ijw} is the individual deviation of the cluster average. The subscripts w and b indicate the *within* and the *between* parts in this decomposition. Similarly

$$X_{ij} = X_{jb} + X_{ijw} \tag{4}$$

where X_{jb} would be the cluster average (such as cluster average SES status) and X_{ijw} is the individual deviation from the cluster average. The observed variables are Y_{ij} and X_{ij} while $Y_{jb}, Y_{ijw}, X_{jb}, X_{ijw}$ are all unobserved. A linear two-level regression models with separate individual and group level effects is defined by the following equations

$$Y_{ijw} = \beta_w X_{ijw} + \varepsilon_{ij} \tag{5}$$

$$Y_{jb} = \mu + \beta_b X_{jb} + \varepsilon_j. \tag{6}$$

Using equation (5) we can substitute Y_{ijw} in equation (3) to get

$$Y_{ij} = Y_{jb} + \beta_w X_{ijw} + \varepsilon_{ij} \tag{7}$$

It is now easy to see that equations (1) and (2) are designed to approximate equations (7) and (6) respectively. Note that the random intercept β_{0j} in equations (1) and (2) has the same interpretation as cluster average Y_{jb} in equations (7) and (6).
The basic assumptions for this model are as follows:

(i) the variables in equation (5) are independent of the variables in equation (6)
(ii) ε_{ij} is a mean 0 residual independent of X_{ijw}
(iii) ε_j is a mean 0 residual independent of X_{jb}

In addition to these assumptions a fourth assumption is frequently made

(iv) all of the variables in equation (5) and (6) are normally distributed.

This fourth assumption however in not really needed. This is because normality is not an assumption that is needed for mean and variance/covariance structure estimation such as the one defined by equation (6) and (5). Central limit theorem guarantees that the ML estimates are consistent even when the variables are non-normal. In addition robust ML estimation (given by Mplus

MLR estimator) produces standard error estimates that are valid even when the variables are non-normal.

The parameters of interest in the two-level regression model are the intercept μ, and the regression coefficients β_w and β_b as well as the residual variance parameters $Var(\varepsilon_{ij}) = \theta_w$ and $Var(\varepsilon_j) = \theta_b$. There are also three auxiliary parameters in the model $\mu_x = E(X_{ij})$, $Var(X_{ijw}) = \psi_w$ and $Var(X_{jb}) = \psi_b$. Given the regression model (5) and (6) the following equations hold

$$\beta_w = \frac{Cov(Y_{ijw}, X_{ijw})}{Var(X_{jw})} \tag{8}$$

$$\beta_b = \frac{Cov(Y_{jb}, X_{jb})}{Var(X_{jb})} \tag{9}$$

In the following sections we will describe three different ways for specifying this model in Mplus and will discussed the merits of the corresponding estimates. The main difference between the three approaches is in how they specify the unobserved covariates X_{ijw} and X_{jb}. We do not provide references for these approaches but our experience is that all three are somewhat frequently used. For simplicity we assume that the sample is balanced, i.e., that there are k clusters all of size l and the total sample size is $n = kl$.

2 Latent Variable Covariates

With this approach we treat X_{ijw} and X_{jb} as latent (unobserved) covariates. This is the default setting in Mplus. In this case the X and Y variables have the within-between status, i.e., the variables X and Y can be used on both levels in the model description. The two-level regression model is specified as

```
%WITHIN%
y ON x;
%BETWEEN%
y ON x;
```

In the within level section y and x refer to Y_{ijw} and X_{ijw} while in the between level section y and x refer to Y_{jb} and X_{jb}. This model specification

corresponds exactly to the one given in equations (5) and (6) and maximum likelihood estimation implemented in Mplus gives consistent estimates.

Note however that if the two-level regression model described here is a part of a bigger model and is estimated simultaneously with other model components that require numerical integration then this specification is not the default. This is done so that the covariate decomposition does not increase the dimensions of numerical integration, which can result in substantially slower model estimation. When a model is estimated with numerical integration, each covariate has to be specified as a within or between only variable. It is possible to specify latent variable covariates within the numerical integration estimation but this is not done automatically because it will increases the computational time.

3 Fixed Covariates

It is possible to estimate the unobserved covariates X_{ijw} and X_{jb} and to use these observed estimates in the regression equations (5) and (6). In general one could hope that the errors in these estimates will cancel out and that the regression coefficients in (5) and (6) will remain consistent, however we will see below that this is not the case. We considerer two different approaches to covariance estimation both of which have been described in Raudenbush & Bryk (2002).

3.1 Group-Mean Centering

Consider the group-mean centering approach described in Raudenbush-Bryk (2002), Table 5.11, page 140. With this approach we estimate X_{jb} by Z_{jb}

$$Z_{jb} = \bar{X}_{.j} = \frac{1}{l}\sum_{i=1}^{l} X_{ij} = X_{jb} + \frac{1}{l}\sum_{i=1}^{l} X_{ijw}$$

and X_{ijw} by Z_{ijw}

$$Z_{ijw} = X_{ij} - \bar{X}_{.j} = X_{ijw} - \frac{1}{l}\sum_{i=1}^{l} X_{ijw}$$

We then use the observed covariates Z_{ijw} and Z_{jb} instead of the unobserved covariates X_{ijw} and X_{jb} in the regression equations (5) and (6). The observed covariate can be constructed in Mplus through the define command

or outside of Mplus, for example in Excel. The two-level regression model is specified by first declaring Z_{ijw} and Z_{jb} as a within-only and between-only covariates via the variable section commands

```
WITHIN=zw;
BETWEEN=zb;
```

The model is then defined by

```
%WITHIN%
y ON zw;
%BETWEEN%
y ON zb;
```

The parameter estimates however obtained this way will be biased. This bias is not due to the Mplus estimator, but it is due to the fact that we replace the unobserved covariates with approximate observed quantities. This approximation is the source of the bias and using alternative multilevel modeling software such as HLM will produce exactly the same bias. The bias on the between level is computed as follows

$$E(\hat{\beta}_b)-\beta_b = \frac{Cov(Y_{ij}, Z_{jb})}{Var(Z_{jb})}-\beta_b = \frac{Cov(Y_{jb}, Z_{jb}) + Cov(Y_{ijw}, Z_{jb})}{Var(Z_{jb})}-\beta_b = \quad (10)$$

$$\frac{\beta_b\psi_b + \beta_w\psi_w/l}{\psi_b + \psi_w/l} - \beta_b = \frac{(\beta_w - \beta_b)\psi_w/l}{\psi_b + \psi_w/l} \quad (11)$$

The bias on the within level is 0 however since

$$E(\hat{\beta}_w) = \frac{Cov(Y_{ijw}, Z_{ijw})}{Var(Z_{ijw})} = \frac{Cov(Y_{ijw}, X_{ijw})}{Var(X_{ijw})} = \beta_w$$

Note that the between level regression bias decreases to 0 when the cluster size l increases to infinity but will not disappear if the sample size n increases while the cluster size l is held constant. Also note that this bias is particularly large when the ICC of the X covariate is small, a rather common situation. Essentially replacing the latent covariates with their observed analogues results in shrinking the between level effect of that covariate. The smaller the covariate's ICC is the bigger the shrinking factor.

3.2 Grand-Mean Centering

Consider the grand-mean centering approach described in Raudenbush-Bryk (2002), Table 5.11, page 140. With this approach we estimate X_{jb} again by Z_{jb} as in the previous section and X_{ijw} by Z_{ij}

$$Z_{ij} = X_{ij} - \bar{X}_{..}$$

where

$$X_{..} = \frac{1}{k}\sum_{j=1}^{k} X_{.j}$$

We then use the observed covariates Z_{ij} and Z_{jb} instead of the unobserved covariates X_{ijw} and X_{jb} in the regression equations (5) and (6). The two-level regression model is specified by first declaring Z_{ij} and Z_{jb} as a within-only and between-only covariates via the variable section commands

```
WITHIN=z;
BETWEEN=zb;
```

The model is then defined by

```
%WITHIN%
y on z;
%BETWEEN%
y on zb;
```

Since assumption (i) is not valid here the parameter estimates will not satisfy equations (8) and (9). However there is a simple algebraic transformation between the model estimated in this section and the model estimated in the previous section.

$$Y_{ij} = Y_{ijw} + Y_{jb} = \mu + \hat{\beta}_w Z_{ij} + \hat{\beta}_b Z_{jb} + \varepsilon_{ij} + \varepsilon_j =$$

$$\mu - \hat{\beta}_w \bar{X}_{..} + \hat{\beta}_w Z_{ijw} + (\hat{\beta}_b + \hat{\beta}_w) Z_{jb} + \varepsilon_{ij} + \varepsilon_j$$

Therefore the bias for the within level regression will be zero while the bias for the between level regression will be

$$\beta_w + \frac{(\beta_w - \beta_b)\psi_w/l}{\psi_b + \psi_w/l}$$

Note that even when l increases to infinity this bias will not disappear. Quantities such as the ICC of the Y variable will be severely biased. Additional problems can arise from the violation of assumption (i). For example using expected information matrix while violating assumption (i) could lead to very poor standard error estimates.

Note also that the grand mean centering in this estimation has only a marginal role. The centering affects only the intercept estimates but doesn't affect the slope estimates. This centering is similar to the grand-mean centering for regular linear regression, which also affects only the intercept estimates but not the slope estimates. Thus if we estimate X_{ijw} by the uncentered X_{ij} instead of the centered Z_{ij} we will obtain the exact same bias in the slope estimates.

4 References

S. Raudenbush & A. Bryk (2002) Hierarchical Linear Models: Applications and Data Analysis Methods. Second Edition. Thousand Oaks. Sage Publications, Inc.