

## Centering

HLM offers the options to use predictors as they are, or to use them after grand- or group-mean centering them. The choice of centering method is dictated by the question studied, and great care should be taken to select a form of centering appropriate to the model considered, as the interpretation of coefficients in the model is dependent on the type of centering used.

We start with an example where a simple linear transformation of a predictor is used to ensure that the interpretation of the average estimated outcome is useful, after which some examples of the different types of centering are given for the same data set.

An intercept term is generally interpreted as the expected average of the dependent variable given that all predictors in a model are equal to zero. This will be the case, for example, in a model where the test score of students nested within schools are studied. A single level-1 predictor IQ is included in the model in an attempt to explain the variability in the test scores of the students, and only the intercept is allowed to vary randomly over the schools (level-2 units).

### Model 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}IQ_{ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

The meaning of the intercept in the Level-1 model depends upon the location of the level-1 predictors included in the model. From the mixed model formulation given below, we see that  $\gamma_{00}$  can be interpreted as the expected test score for a student from group  $j$  with intelligence quotient equal to zero. Although a valid numerical interpretation, a more appropriate choice of the location of IQ will render the results more easily interpretable.

$$Y_{ij} = \gamma_{00} + \gamma_{10}IQ_{ij} + u_{0j} + r_{ij}$$

In the next example, data of a financial nature are studied. Seven consecutive annual measurements from a number of firms on both their annual return on assets (represented here by the variable ROA) and the natural logarithm of their total assets in the same year (represented by the predictor LNASTS) are available.

### Model 2:

$$ROA_{ij} = \beta_{0j} + \beta_{1j}(LNASTS) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

The model can be rewritten as

$$ROA_{ij} = \gamma_{00} + \gamma_{10}(LNASTS) + u_{0j} + (LNASTS)u_{1j} + r_{ij}$$

In this example,  $\gamma_{00}$  can be interpreted as the expected average return on assets for a year (annual measurements being the level-1 units) from group  $j$  with log of assets equal to zero. Also note that the variance-covariance matrix at firm level is of order  $2 \times 2$ , with three non-duplicated elements:  $\text{var}(u_0)$ ,  $\text{var}(u_1)$ ,  $\text{cov}(u_1, u_0)$ .

A number of options concerning the location of predictors can be considered. In some cases, a proper choice of location will be required in order to insure numerical stability in estimating hierarchical linear models. Predictors can, for example, be transformed to deviations from the grand mean or to deviations from the group means. In the case of Model 2, the firm mean of the log of the assets over the years for which measurements are available can be used. However, centering in multilevel models can have different and unexpected results, depending on the way in which the variables are centered.

Two main advantages of centering the predictors are:

- Obtaining estimates of  $\beta_{0j}$  and other effects that are easier to interpret, so that the statistical results can be related to the theoretical concerns that motivate the research.
- Removing high correlations between the random intercept and slopes, and high correlations between first- and second-level variables and cross-level interactions (for a detailed discussion of this aspect, the reader is referred to Kreft and de Leeuw, (1998), pp. 135 to 137).

Model 2 will be used to illustrate some of the options available and the impact of implementation of such options on the interpretability of the coefficients in the model.

For model 2, it follows that

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

where it is assumed that  $\beta_{0j}$  has a normal distribution with zero mean and variance  $\tau_{00}$ . Under this assumption, it follows that the expected value of  $\beta_{0j}$  is  $\gamma_{00}$  for all level-2 units. Hence  $\gamma_{00}$  is the population average value of ROA, given LNASTS = 0. It further follows that the variance of  $\beta_{0j}$ ,  $\tau_{00}$ , is a measure of the variability of the intercept between firms.

An estimate of  $\beta_{0j}$  is obtained from the equation

$$\hat{\beta}_{0j} = \hat{\gamma}_{00} + \hat{u}_{0j}$$

where  $\hat{\gamma}_{00}$  is given in the HLM output under fixed estimates and  $\hat{u}_{0j}$  can be obtained from the residual files. An estimate of  $\tau_{00}$  is given in the random part of the output.  $\hat{u}_{0j}$  is a measure of the amount that the intercept for firm  $j$  differs from the population intercept ( $\hat{\gamma}_{00}$ ) value.

Finally,  $\beta_{1j}$  is the expected LNASTS slope, and  $u_{1j}$  the unique increment in LNASTS slope associated with firm  $j$ .  $\text{Var}(u_{1j})$  is the variation in the log of assets slope over the firms at level-2 of the hierarchy.

In the current situation, both problems that can be addressed by centering are present: the interpretation of  $\beta_{0j}$  in model 2 is problematical, and the output produced for this model contain various warnings indicating the presence of collinearity between the predictors in the model.

### Model 3: Grand mean centering

In the grand mean centered model, the explanatory variable(s) are centered around the overall mean. Continuing with the financial data example given in model 2, the log of assets (LNASTS) used in model 2 is replaced with deviations from  $\overline{\text{LNASTS}}$ , which represents the grand mean of all the log of assets values, irrespective of firm.

#### Level-1 model:

$$ROA_{ij} = \beta_{0j} + (\text{LNASTS}_{ij} - \overline{\text{LNASTS}}) + r_{ij}$$

#### Level-2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

- In this model,  $\beta_{0j}$  is the expected value of  $ROA_{ij}$  given  $\text{LNASTS}_{ij} = \overline{\text{LNASTS}}$ , that is the expected value of a measurement  $i$  from firm  $j$  with log of assets equal to the grand mean of all the log(assets) values. Each  $\beta_{0j}$  is interpreted as the mean outcome for the  $j$ -th level-2 unit adjusted for differences in  $\text{LNASTS}_{ij}$  within this level-2 unit.
- The variance of  $\beta_{0j}$  has a different interpretation, too. It is now the variance between level-2 units in the adjusted means.
- $\gamma_{00}$  is the adjusted mean return on assets/average firm means on ROA across firms.
- $\beta_{1j}$  is the mean log of assets slope, adjusted for differences in LNASTS.
- The variance of  $\beta_{1j}$  is the variation between groups in the mean LNASTS slope.
- The average LNASTS-ROA regression slope across the firms is denoted by  $\gamma_{10}$ .
- The variance at level-1 in  $r_{ij}$  is the residual variance after an adjustment has been made for the level-1 predictor  $\text{LNASTS}_{ij}$ .

It can be shown that the two models considered thus far are equivalent linear models (see Kreft et al. (1995) and Kreft and De Leeuw (1998) pages 106 to 114). Equivalent models will not have the same parameter estimates, but estimates from one can be translated into the estimates from another. They will have the same fit, same predicted values and the same residuals.

### Model 4: Group mean centering

In the group mean centered model, the explanatory variable(s) are centered around the group mean, in this case the firm mean. In the model given below,  $LNASTS_{-j}$  represents the group mean of all the log of assets values from firm  $j$ .

#### Level-1 model:

$$ROA_{ij} = \beta_{0j} + (LNASTS_{ij} - LNASTS_{-j}) + r_{ij}$$

#### Level-2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

- $\beta_{0j}$  is the unadjusted mean for group  $j$ .
- $\gamma_{00}$  is the average of the return on assets across the population of firms. It can also be interpreted as the expected return on assets for a measurement  $i$  from firm  $j$  with log of assets equal to the mean log of assets for firm  $j$ .
- The variance of  $\beta_{0j}$  represents the between group variance in the outcome variable, with  $u_{0j}$  the unique increment to the intercept associated with firm  $j$ .
- $\beta_{1j}$  is the log of assets slope, adjusted for differences in group mean LNASTS.
- The variance of  $\beta_{1j}$  is the variation among groups in the mean LNASTS slope.
- $\gamma_{10}$  is the average LNASTS-ROA regression slope across firms.
- $u_{1j}$  is the unique increment to the LNASTS slope associated with firm  $j$ .
- $r_{ij}$  is the residual variance at level-1 after controlling for measurement LNASTS.

In this model, the between-group variation of the log of assets (LNASTS) is no longer estimated separately. As a result, the level-2 intercept relationships simply represent the group (firm) level relationship between the level-2 predictor and the outcome variable, as can be seen here from the interpretation of  $\gamma_{00}$  given above.

It can also be shown that the group-mean centered model described here is not linearly numerically equivalent to either the raw score or grand mean centered model. The two special cases when the group centered model is equivalent to the other models are described in Kreft & De Leeuw (1998), p 109. In all other cases the lack of linear equivalence implies that the fit of the group centered model, as well as the estimates, are different for this model.

### Comparison of centered models

The results for the raw score, grand mean centered and group mean centered models are given in the table below.

Effect	$\gamma_{00}$ (s.e.)	$\gamma_{:0}$ (s.e.)	$Var(r_{ij})$	$var(u_{0j})$	$Var(u_{1j})$	Deviance, no of parameters
LNASTS (raw score model, uncentered)	13.3978 (1.2123)	-0.9643 (-0.1334)	14.2727	179.7382	1.7322	19 135 6
LNASTS (Grand mean centered)	5.0709 (0.1989)	-0.9626 (0.1339)	14.2623	14.3438	1.7637	19 135 6
LNASTS (Group mean centered)	5.1415 (0.2145)	0.3270 (0.3232)	12.8307	20.0190	16.2573	19 113 6

From the table, we see that

- The estimate of  $\gamma_{00}$  in the case of the raw score model is 13.3978. Recall that this is the expected return of assets for a measurement with the log of assets, LNASTS, equal to 0. The level-2 residual associated with  $\gamma_{00}$  has a variance of 179.7382, which is the estimated variation in intercepts over the firms.
- For the grand mean centered model,  $\gamma_{00}$  is estimated at 5.0709. This is the adjusted mean return on assets.  $var(u_{0j})$  is estimated at 14.3438, and represents the variance among the groups in the adjusted means.
- In the case of the group mean centered model,  $\gamma_{00}$  is the average of the return on assets across the population of firms. Here  $var(u_{0j})$  is estimated as 20.0190, with  $u_{0j}$  the unique increment to the intercept associated with firm.

In the case of  $\gamma_{:0}$ , the estimates obtained from the raw score and grand mean centered models are essentially the same. The estimate from the group centered model, however, is smaller.

- In the grand mean centered model,  $\gamma_{:0}$  is the pooled-within regression coefficient for LNASTS.
- In the raw score model,  $\gamma_{:0}$  is the main effect for the LNASTS slope.
- In the group mean centered model, however,  $\gamma_{:0}$  is the average LNASTS-ROA regression slope across firms.

Turning to the variance components, we find that  $Var(r_{ij})$  and  $Var(u_{1j})$  were the same for the raw score and grand mean centered models, but different for the group mean centered model. In the case of the group mean centered model, the variance at level-1 in  $r_{ij}$  is now the residual variance after controlling for the level-1 predictor  $LNASTS_{ij}$  within each group. As group-mean centering individual measurements leads to smaller differences between the original measurements and the group mean than would be obtained when differences from the grand mean is calculated, a reduced level-1 variance component is not unexpected. In the table below, a small illustration of the effect of group mean and grand mean centering is given for 2 hypothetical level-2 units.

Unit 1			Unit 2		
$x_{ij}$	$x_{ij} - x_{-j}$	$x_{ij} - \bar{x}$	$x_{ij}$	$x_{ij} - x_{-j}$	$x_{ij} - \bar{x}$
1	-1	-10	19	-1	8
2	0	-9	20	0	9
3	1	-8	21	1	10
Total: 6	Total: 0	Total: -27	Total: 60	Total: 0	Total: 27

As centering reduces the correlation between, in this example, the intercept and the LNASTS slope at level-2,  $\text{var}(u_{0j})$  decreases dramatically.

From the deviance statistics for these models, the group-mean centered model produced the best fit to the data. The between-groups variance of the log of assets is, however, not estimated in this model. Thus an uncorrected between-groups effect is measured, which is not adjusted for the mean effect of the centered level-1 predictor. By adding the means of this predictor to the level-2 equation for the intercept, the correct between-group variance is again estimated. This becomes clear when the following two mixed-model formulations of the group mean centered models are compared:

**Without the means reintroduced:**

$$\begin{aligned}
 ROA_{ij} &= \gamma_{00} + \gamma_{10} (LNASTS_{ij} - LNASTS_{-j}) + \text{random components} \\
 &= \gamma_{00} + \gamma_{10} (LNASTS_{ij}) - \gamma_{10} (LNASTS_{-j}) + \text{random components}
 \end{aligned}$$

**With means introduced in the intercept equation:**

$$\begin{aligned}
 ROA_{ij} &= \gamma_{00} + \gamma_{01} (LNASTS_{-j}) + \gamma_{10} (LNASTS_{ij} - LNASTS_{-j}) + \text{random components} \\
 &= \gamma_{00} + \gamma_{01} (LNASTS_{-j}) + \gamma_{10} (LNASTS_{ij}) - \gamma_{10} (LNASTS_{-j}) + \text{random components}
 \end{aligned}$$

As is clear from this illustration, the type of centering used has different consequences. A researcher should consider these options carefully, and refrain from centering simply to improve statistical stability. If group-mean centering is used, it is advisable to add the group means for the centered predictor to the model as well, as illustrated above in order to avoid estimating an uncorrected between-groups effect.

**Further reading:**

- Bryk, A.S. & Raudenbush, S.W., *Hierarchical Linear Models*, Sage, (1992): General centering information is given on pages 25 to 28 of this text. A group mean centered model is discussed in detail in Chapter 2, while a grand mean centered example can be found in Chapter 4.
- Kreft, I. & De Leeuw, J., *Introducing Multilevel Modeling*, Sage, (1998): A detailed discussion of the consequences of centering, with examples of the different options using one data set can be found on pages 106 to 114.
- Snijders, T. & Bosker, R., Sage, (2000): pp. 80-81 and pp. 52-54 of this text covers the topic of centering and provides a simple, yet clear example of the use of a level-1 predictor and its mean as covariate at level-2 of the hierarchy.

- Hofmann, D.A. & Gavin, Mark B., Centering Decisions in Hierarchical Linear Models: Implications for Research in Organizations, *Journal of Management*, 1998, Vol 24, No 5, pages 623-641.