

CHAPTER 9

EXAMPLES: MULTILEVEL MODELING WITH COMPLEX SURVEY DATA

Complex survey data refers to data obtained by stratification, cluster sampling and/or sampling with an unequal probability of selection. Complex survey data are also referred to as multilevel or hierarchical data. For an overview, see Muthén and Satorra (1995). There are two approaches to the analysis of complex survey data in Mplus.

One approach is to compute standard errors and a chi-square test of model fit taking into account stratification, non-independence of observations due to cluster sampling, and/or unequal probability of selection. Subpopulation analysis is also available. With sampling weights, parameters are estimated by maximizing a weighted loglikelihood function. Standard error computations use a sandwich estimator. This approach can be obtained by specifying `TYPE=COMPLEX` in the `ANALYSIS` command in conjunction with the `STRATIFICATION`, `CLUSTER`, `WEIGHT`, and/or `SUBPOPULATION` options of the `VARIABLE` command. Observed outcome variables can be continuous, censored, binary, ordered categorical (ordinal), unordered categorical (nominal), counts, or combinations of these variable types. The implementation of these methods in Mplus is discussed in Asparouhov (2005, 2006) and Asparouhov and Muthén (2005, 2006a).

A second approach is to specify a model for each level of the multilevel data thereby modeling the non-independence of observations due to cluster sampling. This is commonly referred to as multilevel modeling. The use of sampling weights in the estimation of parameters, standard errors, and the chi-square test of model fit is allowed. Both individual-level and cluster-level weights can be used. With sampling weights, parameters are estimated by maximizing a weighted loglikelihood function. Standard error computations use a sandwich estimator. This approach can be obtained by specifying `TYPE=TWOLEVEL` in the `ANALYSIS` command in conjunction with the `CLUSTER`, `WEIGHT`, `WTSCALE`, `BWEIGHT`, and/or `BWTSCALE` options of the

CHAPTER 9

VARIABLE command. Observed outcome variables can be continuous, censored, binary, ordered categorical (ordinal), unordered categorical (nominal), counts, or combinations of these variable types. The examples in this chapter illustrate this approach.

The two approaches described above can be combined by specifying `TYPE=COMPLEX TWOLEVEL` in the ANALYSIS command in conjunction with the STRATIFICATION, CLUSTER, WEIGHT, WTSCALE, BWEIGHT, and BWTSCALE options of the VARIABLE command. When there is clustering due to both primary and secondary sampling stages, the standard errors and chi-square test of model fit are computed taking into account the clustering due to the primary sampling stage using `TYPE=COMPLEX` whereas clustering due to the secondary sampling stage is modeled using `TYPE=TWOLEVEL`.

A distinction can be made between cross-sectional data in which non-independence arises because of cluster sampling and longitudinal data in which non-independence arises because of repeated measures of the same individuals across time. With cross-sectional data, the number of levels in Mplus is the same as the number of levels in conventional multilevel modeling programs. Mplus allows two-level modeling. With longitudinal data, the number of levels in Mplus is one less than the number of levels in conventional multilevel modeling programs because Mplus takes a multivariate approach to repeated measures analysis. Longitudinal models are two-level models in conventional multilevel programs, whereas they are single-level models in Mplus. These models are discussed in Chapter 6. Three-level analysis where time is the first level, individual is the second level, and cluster is the third level is handled by two-level modeling in Mplus (see also Muthén, 1997).

The general latent variable modeling framework of Mplus allows the integration of random effects and other continuous latent variables within a single analysis model. Random effects are allowed for both independent and dependent variables and both observed and latent variables. Random effects representing across-cluster variation in intercepts and slopes or individual differences in growth can be combined with factors measured by multiple indicators on both the individual and cluster levels. In line with SEM, regressions among random effects, among factors, and between random effects and factors are allowed.

Examples: Multilevel Modeling With Complex Survey Data

Multilevel models can include regression analysis, path analysis, confirmatory factor analysis (CFA), item response theory (IRT) analysis, structural equation modeling (SEM), latent class analysis (LCA), latent transition analysis (LTA), latent class growth analysis (LCGA), growth mixture modeling (GMM), discrete-time survival analysis, continuous-time survival analysis, and combinations of these models.

Two-level modeling in Mplus has three estimator options. The first estimator option is full-information maximum likelihood which allows continuous, censored, binary, ordered categorical (ordinal), unordered categorical (nominal), counts, or combinations of these variable types; random intercepts and slopes; and missing data. With longitudinal data, maximum likelihood estimation allows modeling of individually-varying times of observation and random slopes for time-varying covariates. Non-normality robust standard errors and a chi-square test of model fit are available. The second estimator option is limited-information weighted least squares (Asparouhov & Muthén, 2007) which allows continuous, binary, ordered categorical (ordinal), and combinations of these variables types; random intercepts; and missing data. The third estimator option is the Muthén limited information estimator (MUML; Muthén, 1994) which is restricted to models with continuous outcomes, random intercepts, and no missing data.

All multilevel models can be estimated using the following special features:

- Single or multiple group analysis
- Missing data
- Complex survey data
- Latent variable interactions and non-linear factor analysis using maximum likelihood
- Random slopes
- Individually-varying times of observations
- Linear and non-linear parameter constraints
- Indirect effects including specific paths
- Maximum likelihood estimation for all outcome types
- Wald chi-square test of parameter equalities

For continuous, censored with weighted least squares estimation, binary, and ordered categorical (ordinal) outcomes, multiple group analysis is specified by using the GROUPING option of the VARIABLE command

for individual data or the `NGROUPS` option of the `DATA` command for summary data. For censored with maximum likelihood estimation, unordered categorical (nominal), and count outcomes, multiple group analysis is specified using the `KNOWNCLASS` option of the `VARIABLE` command in conjunction with the `TYPE=MIXTURE` option of the `ANALYSIS` command. The default is to estimate the model under missing data theory using all available data. The `LISTWISE` option of the `DATA` command can be used to delete all observations from the analysis that have missing values on one or more of the analysis variables. Corrections to the standard errors and chi-square test of model fit that take into account stratification, non-independence of observations, and unequal probability of selection are obtained by using the `TYPE=COMPLEX` option of the `ANALYSIS` command in conjunction with the `STRATIFICATION`, `CLUSTER`, and `WEIGHT` options of the `VARIABLE` command. Latent variable interactions are specified by using the `|` symbol of the `MODEL` command in conjunction with the `XWITH` option of the `MODEL` command. Random slopes are specified by using the `|` symbol of the `MODEL` command in conjunction with the `ON` option of the `MODEL` command. Individually-varying times of observations are specified by using the `|` symbol of the `MODEL` command in conjunction with the `AT` option of the `MODEL` command and the `TSCORES` option of the `VARIABLE` command. Linear and non-linear parameter constraints are specified by using the `MODEL CONSTRAINT` command. Indirect effects are specified by using the `MODEL INDIRECT` command. Maximum likelihood estimation is specified by using the `ESTIMATOR` option of the `ANALYSIS` command. The `MODEL TEST` command is used to test linear restrictions on the parameters in the `MODEL` and `MODEL CONSTRAINT` commands using the Wald chi-square test.

Graphical displays of observed data and analysis results can be obtained using the `PLOT` command in conjunction with a post-processing graphics module. The `PLOT` command provides histograms, scatterplots, plots of individual observed and estimated values, and plots of sample and estimated means and proportions/probabilities. These are available for the total sample, by group, by class, and adjusted for covariates. The `PLOT` command includes a display showing a set of descriptive statistics for each variable. The graphical displays can be edited and exported as a `DIB`, `EMF`, or `JPEG` file. In addition, the data for each graphical display can be saved in an external file for use by another graphics program.

Examples: Multilevel Modeling With Complex Survey Data

Following is the set of cross-sectional multilevel modeling examples included in this chapter:

- 9.1: Two-level regression analysis for a continuous dependent variable with a random intercept
- 9.2: Two-level regression analysis for a continuous dependent variable with a random slope
- 9.3: Two-level path analysis with a continuous and a categorical dependent variable*
- 9.4: Two-level path analysis with a continuous, a categorical, and a cluster-level observed dependent variable
- 9.5: Two-level path analysis with continuous dependent variables and random slopes*
- 9.6: Two-level CFA with continuous factor indicators and covariates
- 9.7: Two-level CFA with categorical factor indicators and covariates*
- 9.8: Two-level CFA with continuous factor indicators, covariates, and random slopes
- 9.9: Two-level SEM with categorical factor indicators on the within level and cluster-level continuous observed and random intercept factor indicators on the between level
- 9.10: Two-level SEM with continuous factor indicators and a random slope for a factor*
- 9.11: Two-level multiple group CFA with continuous factor indicators

Following is the set of longitudinal multilevel modeling examples included in this chapter:

- 9.12: Two-level growth model for a continuous outcome (three-level analysis)
- 9.13: Two-level growth model for a categorical outcome (three-level analysis)*
- 9.14: Two-level growth model for a continuous outcome (three-level analysis) with variation on both the within and between levels for a random slope of a time-varying covariate*
- 9.15: Two-level multiple indicator growth model with categorical outcomes (three-level analysis)

CHAPTER 9

- 9.16: Linear growth model for a continuous outcome with time-invariant and time-varying covariates carried out as a two-level growth model using the DATA WIDETOLONG command
- 9.17: Two-level growth model for a count outcome using a zero-inflated Poisson model (three-level analysis)*
- 9.18: Two-level continuous-time survival analysis using Cox regression with a random intercept

* Example uses numerical integration in the estimation of the model. This can be computationally demanding depending on the size of the problem.

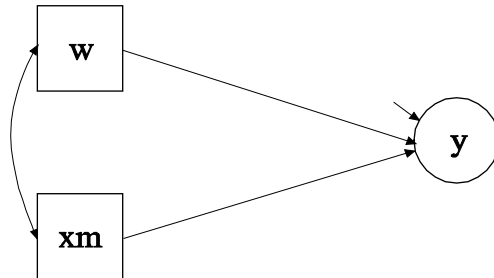
EXAMPLE 9.1: TWO-LEVEL REGRESSION ANALYSIS FOR A CONTINUOUS DEPENDENT VARIABLE WITH A RANDOM INTERCEPT

```
TITLE:      this is an example of a two-level
             regression analysis for a continuous
             dependent variable with a random intercept
             and an observed covariate
DATA:       FILE = ex9.1a.dat;
VARIABLE:   NAMES = y x w xm clus;
             WITHIN = x;
             BETWEEN = w xm;
             CLUSTER = clus;
             CENTERING = GRANDMEAN (x);
ANALYSIS:   TYPE = TWOLEVEL;
MODEL:      %WITHIN%
             y ON x;
             %BETWEEN%
             y ON w xm;
```

Examples: Multilevel Modeling With Complex Survey Data



Within



Between

In this example, the two-level regression model shown in the picture above is estimated. The dependent variable y in this regression is continuous. Two ways of treating the covariate x are described. In this part of the example, the covariate x is treated as an observed variable in line with conventional multilevel regression modeling. In the second part of the example, the covariate x is decomposed into two latent variable parts.

The within part of the model describes the regression of y on an observed covariate x where the intercept is a random effect that varies across the clusters. In the within part of the model, the filled circle at the end of the arrow from x to y represents a random intercept that is referred to as y in the between part of the model. In the between part of the model, the random intercept is shown in a circle because it is a continuous latent variable that varies across clusters. The between part of the model describes the linear regression of the random intercept y on observed cluster-level covariates w and xm . The observed cluster-level covariate xm takes the value of the mean of x for each cluster. The within and between parts of the model correspond to level 1 and level 2 of a conventional multilevel regression model with a random intercept.

CHAPTER 9

```
TITLE:      this is an example of a two-level
            regression analysis for a continuous
            dependent variable with a random intercept
            and an observed covariate
```

The `TITLE` command is used to provide a title for the analysis. The title is printed in the output just before the Summary of Analysis.

```
DATA:      FILE = ex9.1a.dat;
```

The `DATA` command is used to provide information about the data set to be analyzed. The `FILE` option is used to specify the name of the file that contains the data to be analyzed, `ex9.1a.dat`. Because the data set is in free format, the default, a `FORMAT` statement is not required.

```
VARIABLE:  NAMES = y x w xm clus;
            WITHIN = x;
            BETWEEN = w xm;
            CLUSTER = clus;
            CENTERING = GRANDMEAN (x);
```

The `VARIABLE` command is used to provide information about the variables in the data set to be analyzed. The `NAMES` option is used to assign names to the variables in the data set. The data set in this example contains five variables: `y`, `x`, `w`, `xm`, and `clus`.

The `WITHIN` option is used to identify the variables in the data set that are measured on the individual level and modeled only on the within level. They are specified to have no variance in the between part of the model. The `BETWEEN` option is used to identify the variables in the data set that are measured on the cluster level and modeled only on the between level. Variables not mentioned on the `WITHIN` or the `BETWEEN` statements are measured on the individual level and can be modeled on both the within and between levels. Because `y` is not mentioned on the `WITHIN` statement, it is modeled on both the within and between levels. On the between level, it is a random intercept. The `CLUSTER` option is used to identify the variable that contains clustering information. The `CENTERING` option is used to specify the type of centering to be used in an analysis and the variables that are to be centered. In this example, grand-mean centering is chosen.

Examples: Multilevel Modeling With Complex Survey Data

```
ANALYSIS: TYPE = TWOLEVEL;
```

The ANALYSIS command is used to describe the technical details of the analysis. By selecting TWOLEVEL, a multilevel model with random intercepts will be estimated.

```
MODEL:
      %WITHIN%
      y ON x;
      %BETWEEN%
      y ON w xm;
```

The MODEL command is used to describe the model to be estimated. In multilevel models, a model is specified for both the within and between parts of the model. In the within part of the model, the ON statement describes the linear regression of y on the observed individual-level covariate x . The within-level residual variance in the regression of y on x is estimated as the default.

In the between part of the model, the ON statement describes the linear regression of the random intercept y on the observed cluster-level covariates w and xm . The intercept and residual variance of y are estimated as the default. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator.

CHAPTER 9

Following is the second part of the example where the covariate x is decomposed into two latent variable parts.

```
TITLE:      this is an example of a two-level
             regression analysis for a continuous
             dependent variable with a random intercept
             and a latent covariate
DATA:       FILE = ex9.1b.dat;
VARIABLE:   NAMES = y x w clus;
             BETWEEN = w;
             CLUSTER = clus;
             CENTERING = GRANDMEAN (x);
ANALYSIS:   TYPE = TWOLEVEL;
MODEL:
             %WITHIN%
             y ON x (gamma10);
             %BETWEEN%
             y ON w
             x (gamma01);
MODEL CONSTRAINT:
             NEW(betac);
             betac = gamma01 - gamma10;
```

The difference between this part of the example and the first part is that the covariate x is decomposed into two latent variable parts instead of being treated as an observed variable as in conventional multilevel regression modeling. The decomposition occurs when the covariate x is not mentioned on the WITHIN statement and is therefore modeled on both the within and between levels. When a covariate is not mentioned on the WITHIN statement, it is decomposed into two uncorrelated latent variables,

$$x_{ij} = x_{wij} + x_{bj},$$

where i represents individual, j represents cluster, x_{wij} is the latent variable covariate used on the within level, and x_{bj} is the latent variable covariate used on the between level. This model is described in Muthén (1989, 1990, 1994). The latent variable covariate x_b is not used in conventional multilevel analysis. Using a latent covariate may, however, be advantageous when the observed cluster-mean covariate x_m does not have sufficient reliability resulting in biased estimation of the between-level slope (Asparouhov & Muthén, 2006b; Ludtke et al., 2007).

The decomposition can be expressed as,

$$X_{wij} = X_{ij} - X_{bj},$$

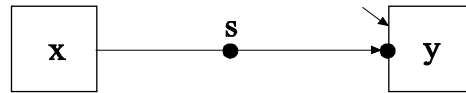
which can be viewed as an implicit, latent group-mean centering of the latent within-level covariate. To obtain results that are not group-mean centered, a linear transformation of the within and between slopes can be done as described below using the MODEL CONSTRAINT command.

In the MODEL command, the label gamma10 in the within part of the model and the label gamma01 in the between part of the model are assigned to the regression coefficients in the linear regression of y on x in both parts of the model for use in the MODEL CONSTRAINT command. The MODEL CONSTRAINT command is used to define linear and non-linear constraints on the parameters in the model. In the MODEL CONSTRAINT command, the NEW option is used to introduce a new parameter that is not part of the MODEL command. This parameter is called betac and is defined as the difference between gamma01 and gamma10. It corresponds to a “contextual effect” as described in Raudenbush and Bryk (2002, p. 140, Table 5.11).

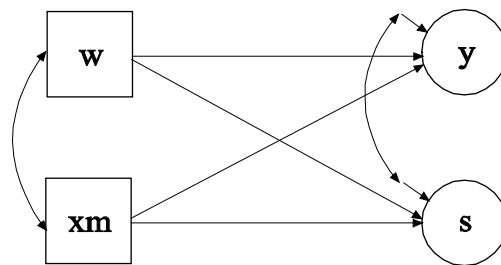
EXAMPLE 9.2: TWO-LEVEL REGRESSION ANALYSIS FOR A CONTINUOUS DEPENDENT VARIABLE WITH A RANDOM SLOPE

```
TITLE:      this is an example of a two-level
             regression analysis for a continuous
             dependent variable with a random slope and
             an observed covariate
DATA:      FILE = ex9.2a.dat;
VARIABLE:  NAMES = y x w xm clus;
             WITHIN = x;
             BETWEEN = w xm;
             CLUSTER = clus;
             CENTERING = GRANDMEAN (x);
ANALYSIS:  TYPE = TWOLEVEL RANDOM;
MODEL:
             %WITHIN%
             s | y ON x;
             %BETWEEN%
             y s ON w xm;
             y WITH s;
```

CHAPTER 9



Within



Between

The difference between this example and the first part of Example 9.1 is that the model has both a random intercept and a random slope. In the within part of the model, the filled circle at the end of the arrow from x to y represents a random intercept that is referred to as y in the between part of the model. The filled circle on the arrow from x to y represents a random slope that is referred to as s in the between part of the model. In the between part of the model, the random intercept and random slope are shown in circles because they are continuous latent variables that vary across clusters. The observed cluster-level covariate xm takes the value of the mean of x for each cluster. The within and between parts of the model correspond to level 1 and level 2 of a conventional multilevel regression model with a random intercept and a random slope.

In the within part of the model, the $|$ symbol is used in conjunction with `TYPE=RANDOM` to name and define the random slope variables in the model. The name on the left-hand side of the $|$ symbol names the random slope variable. The statement on the right-hand side of the $|$ symbol defines the random slope variable. Random slopes are defined using the `ON` option. The random slope s is defined by the linear regression of the dependent variable y on the observed individual-level

Examples: Multilevel Modeling With Complex Survey Data

covariate x . The within-level residual variance in the regression of y on x is estimated as the default.

In the between part of the model, the ON statement describes the linear regressions of the random intercept y and the random slope s on the observed cluster-level covariates w and x_m . The intercepts and residual variances of s and y are estimated as the default. The residuals are correlated as the default. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. An explanation of the other commands can be found in Example 9.1.

Following is the second part of the example that shows an alternative treatment of the observed covariate x .

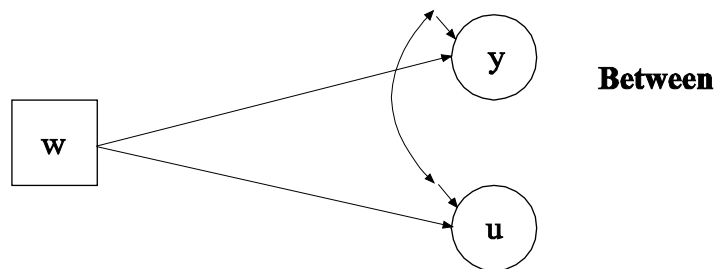
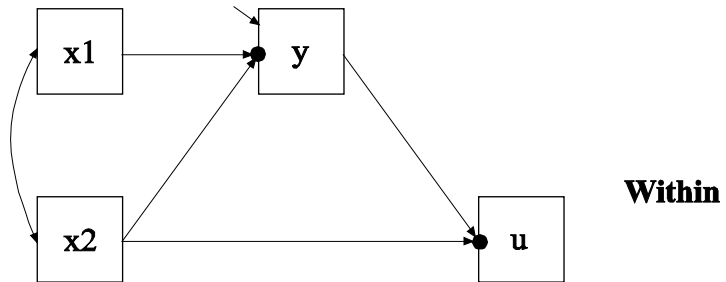
```
TITLE:      this is an example of a two-level
            regression analysis for a continuous
            dependent variable with a random slope and
            a latent covariate
DATA:       FILE = ex9.2b.dat;
VARIABLE:   NAMES = y x w clus;
            BETWEEN = w;
            CLUSTER = clus;
ANALYSIS:   TYPE = TWOLEVEL RANDOM;
MODEL:
            %WITHIN%
            s | y ON x;
            %BETWEEN%
            y s ON w x;
            y WITH s;
```

The difference between this part of the example and the first part of the example is that the covariate x is latent instead of observed on the between level. This is achieved when the individual-level observed covariate is modeled in both the within and between parts of the model. This is requested by not mentioning the observed covariate x on the WITHIN statement in the VARIABLE command. When a random slope is estimated, the observed covariate x is used on the within level and the latent variable covariate x_{bj} is used on the between level. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. An explanation of the other commands can be found in Example 9.1.

EXAMPLE 9.3: TWO-LEVEL PATH ANALYSIS WITH A CONTINUOUS AND A CATEGORICAL DEPENDENT VARIABLE

```

TITLE:      this is an example of a two-level path
            analysis with a continuous and a
            categorical dependent variable
DATA:      FILE IS ex9.3.dat;
VARIABLE:  NAMES ARE u y x1 x2 w clus;
            CATEGORICAL = u;
            WITHIN = x1 x2;
            BETWEEN = w;
            CLUSTER IS clus;
ANALYSIS:  TYPE = TWOLEVEL;
            ALGORITHM = INTEGRATION;
MODEL:
            %WITHIN%
            y ON x1 x2;
            u ON y x2;
            %BETWEEN%
            y u ON w;
OUTPUT:    TECH1 TECH8;
    
```



Examples: Multilevel Modeling With Complex Survey Data

In this example, the two-level path analysis model shown in the picture above is estimated. The mediating variable y is a continuous variable and the dependent variable u is a binary or ordered categorical variable. The within part of the model describes the linear regression of y on x_1 and x_2 and the logistic regression of u on y and x_2 where the intercepts in the two regressions are random effects that vary across the clusters and the slopes are fixed effects that do not vary across the clusters. In the within part of the model, the filled circles at the end of the arrows from x_1 to y and x_2 to u represent random intercepts that are referred to as y and u in the between part of the model. In the between part of the model, the random intercepts are shown in circles because they are continuous latent variables that vary across clusters. The between part of the model describes the linear regressions of the random intercepts y and u on a cluster-level covariate w .

The `CATEGORICAL` option is used to specify which dependent variables are treated as binary or ordered categorical (ordinal) variables in the model and its estimation. The program determines the number of categories of u . The dependent variable u could alternatively be an unordered categorical (nominal) variable. The `NOMINAL` option is used and a multinomial logistic regression is estimated.

In the within part of the model, the first `ON` statement describes the linear regression of y on the individual-level covariates x_1 and x_2 and the second `ON` statement describes the logistic regression of u on the mediating variable y and the individual-level covariate x_2 . The slopes in these regressions are fixed effects that do not vary across the clusters. The residual variance in the linear regression of y on x_1 and x_2 is estimated as the default. There is no residual variance to be estimated in the logistic regression of u on y and x_2 because u is a binary or ordered categorical variable. In the between part of the model, the `ON` statement describes the linear regressions of the random intercepts y and u on the cluster-level covariate w . The intercept and residual variance of y and u are estimated as the default. The residual covariance between y and u is free to be estimated as the default.

By specifying `ALGORITHM=INTEGRATION`, a maximum likelihood estimator with robust standard errors using a numerical integration algorithm will be used. Note that numerical integration becomes increasingly more computationally demanding as the number of factors and the sample size increase. In this example, two dimensions of

integration are used with a total of 225 integration points. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. The OUTPUT command is used to request additional output not included as the default. The TECH1 option is used to request the arrays containing parameter specifications and starting values for all free parameters in the model. The TECH8 option is used to request that the optimization history in estimating the model be printed in the output. TECH8 is printed to the screen during the computations as the default. TECH8 screen printing is useful for determining how long the analysis takes. An explanation of the other commands can be found in Example 9.1.

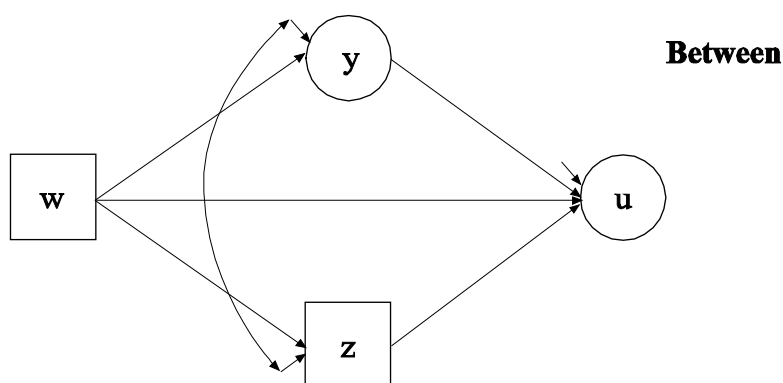
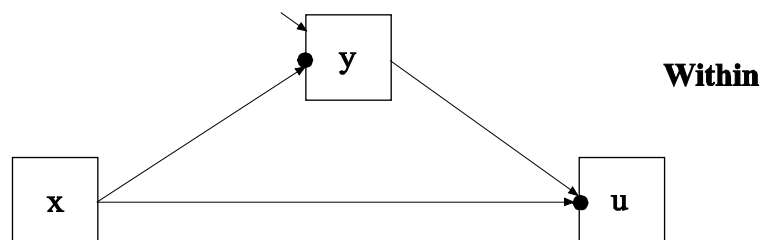
EXAMPLE 9.4: TWO-LEVEL PATH ANALYSIS WITH A CONTINUOUS, A CATEGORICAL, AND A CLUSTER-LEVEL OBSERVED DEPENDENT VARIABLE

```

TITLE:      this is an example of a two-level path
            analysis with a continuous, a categorical,
            and a cluster-level observed dependent
            variable
DATA:      FILE = ex9.4.dat;
VARIABLE:  NAMES ARE u z y x w clus;
            CATEGORICAL = u;
            WITHIN = x;
            BETWEEN = w z;
            CLUSTER = clus;
ANALYSIS:  TYPE = TWOLEVEL;
            ESTIMATOR = WLSM;
MODEL:
            %WITHIN%
            u ON y x;
            y ON x;
            %BETWEEN%
            u ON w y z;
            y ON w;
            z ON w;
            y WITH z;
OUTPUT:    TECH1;

```

Examples: Multilevel Modeling With Complex Survey Data



The difference between this example and Example 9.3 is that the between part of the model has an observed cluster-level mediating variable z and a latent mediating variable y that is a random intercept. The model is estimated using weighted least squares estimation instead of maximum likelihood.

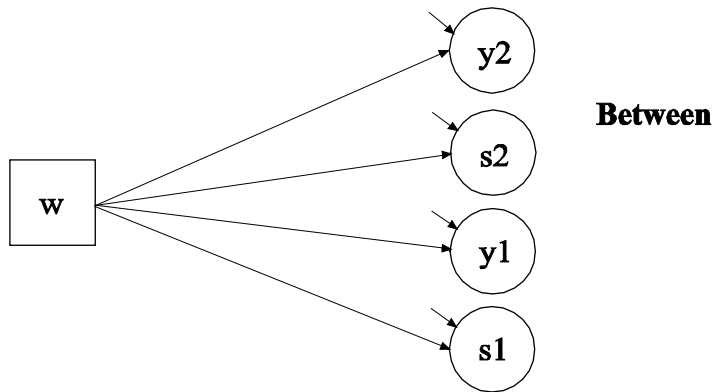
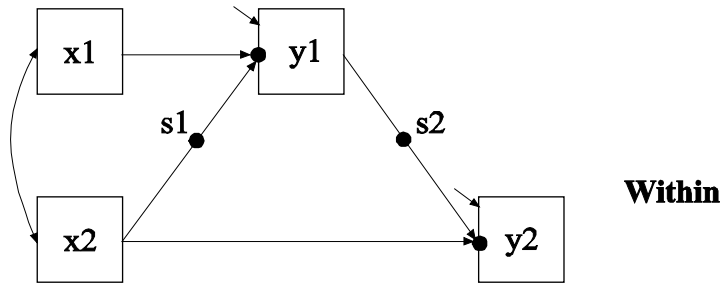
By specifying `ESTIMATOR=WLSM`, a robust weighted least squares estimator using a diagonal weight matrix is used (Asparouhov & Muthén, 2007). The `ESTIMATOR` option of the `ANALYSIS` command can be used to select a different estimator.

In the between part of the model, the first `ON` statement describes the linear regression of the random intercept u on the cluster-level covariate w , the random intercept y , and the observed cluster-level mediating variable z . The third `ON` statement describes the linear regression of the observed cluster-level mediating variable z on the cluster-level covariate w . An explanation of the other commands can be found in Examples 9.1 and 9.3.

EXAMPLE 9.5: TWO-LEVEL PATH ANALYSIS WITH CONTINUOUS DEPENDENT VARIABLES AND RANDOM SLOPES

```
TITLE:      this is an example of two-level path
            analysis with continuous dependent
            variables and random slopes
DATA:       FILE IS ex9.5.dat;
VARIABLE:   NAMES ARE y1 y2 x1 x2 w clus;
            WITHIN = x1 x2;
            BETWEEN = w;
            CLUSTER IS clus;
ANALYSIS:   TYPE = TWOLEVEL RANDOM;
MODEL:
            %WITHIN%
            s2 | y2 ON y1;
            y2 ON x2;
            s1 | y1 ON x2;
            y1 ON x1;
            %BETWEEN%
            y1 y2 s1 s2 ON w;
OUTPUT:    TECH1 TECH8;
```

Examples: Multilevel Modeling With Complex Survey Data



The difference between this example and Example 9.3 is that the model includes two random intercepts and two random slopes instead of two random intercepts and two fixed slopes and the dependent variable is continuous. In the within part of the model, the filled circle on the arrow from the covariate x_2 to the mediating variable y_1 represents a random slope and is referred to as s_1 in the between part of the model. The filled circle on the arrow from the mediating variable y_1 to the dependent variable y_2 represents a random slope and is referred to as s_2 in the between part of the model. In the between part of the model, the random slopes s_1 and s_2 are shown in circles because they are continuous latent variables that vary across clusters.

In the within part of the model, the $|$ symbol is used in conjunction with `TYPE=RANDOM` to name and define the random slope variables in the model. The name on the left-hand side of the $|$ symbol names the random slope variable. The statement on the right-hand side of the $|$ symbol defines the random slope variable. Random slopes are defined

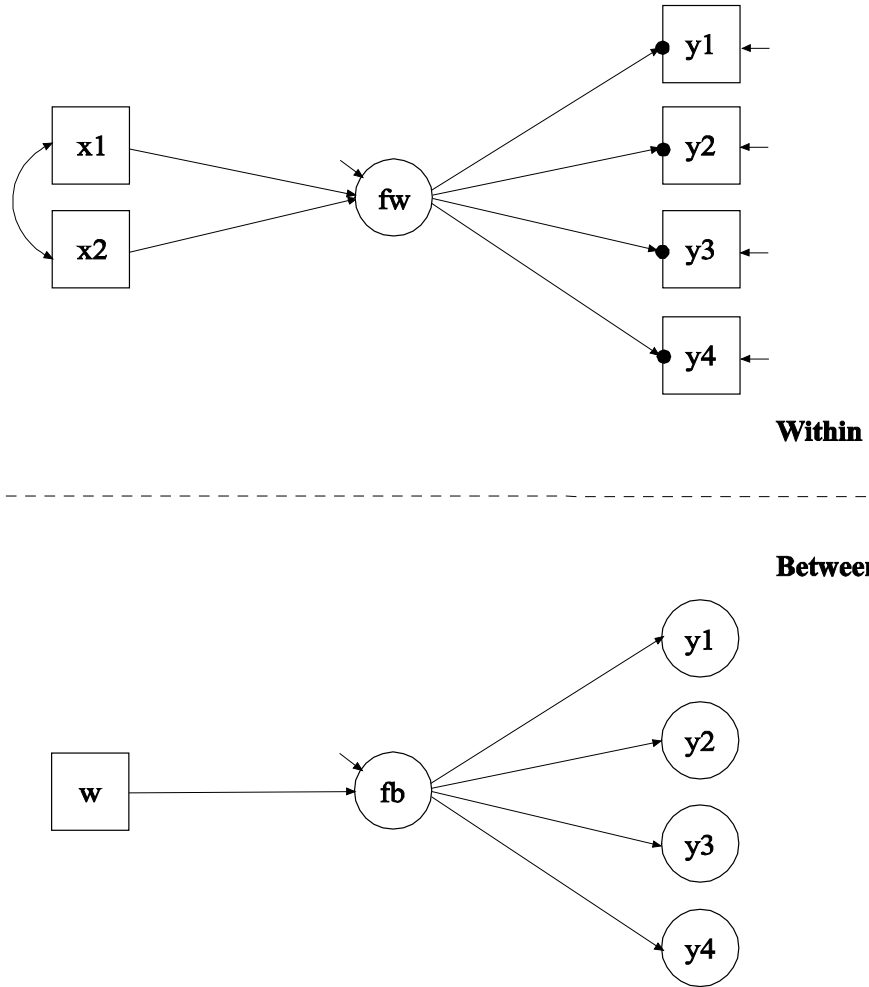
using the ON option. In the first | statement, the random slope s2 is defined by the linear regression of the dependent variable y2 on the mediating variable y1. In the second | statement, the random slope s1 is defined by the linear regression of the mediating variable y1 on the individual-level covariate x2. The within-level residual variances of y1 and y2 are estimated as the default. The first ON statement describes the linear regression of the dependent variable y2 on the individual-level covariate x2. The second ON statement describes the linear regression of the mediating variable y1 on the individual-level covariate x1.

In the between part of the model, the ON statement describes the linear regressions of the random intercepts y1 and y2 and the random slopes s1 and s2 on the cluster-level covariate w. The intercepts and residual variances of y1, y2, s2, and s1 are estimated as the default. The residual covariances between y1, y2, s2, and s1 are fixed at zero as the default. This default can be overridden. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. An explanation of the other commands can be found in Examples 9.1 and 9.3.

EXAMPLE 9.6: TWO-LEVEL CFA WITH CONTINUOUS FACTOR INDICATORS AND COVARIATES

```
TITLE:      this is an example of a two-level CFA with
             continuous factor indicators and
             covariates
DATA:      FILE IS ex9.6.dat;
VARIABLE:  NAMES ARE y1-y4 x1 x2 w clus;
             WITHIN = x1 x2;
             BETWEEN = w;
             CLUSTER = clus;
ANALYSIS:  TYPE = TWOLEVEL;
MODEL:
             %WITHIN%
             fw BY y1-y4;
             fw ON x1 x2;
             %BETWEEN%
             fb BY y1-y4;
             y1-y4@0;
             fb ON w;
```

Examples: Multilevel Modeling With Complex Survey Data



In this example, the two-level CFA model with continuous factor indicators, a between factor, and covariates shown in the picture above is estimated. In the within part of the model, the filled circles at the end of the arrows from the within factor fw to $y1$, $y2$, $y3$, and $y4$ represent random intercepts that are referred to as $y1$, $y2$, $y3$, and $y4$ in the between part of the model. In the between part of the model, the random intercepts are shown in circles because they are continuous latent variables that vary across clusters. They are indicators of the between factor fb . In this model, the residual variances for the factor indicators in the between part of the model are fixed at zero. If factor loadings are

CHAPTER 9

constrained to be equal across the within and the between levels, this implies a model where the regression of the within factor on x_1 and x_2 has a random intercept varying across the clusters.

In the within part of the model, the BY statement specifies that fw is measured by y_1 , y_2 , y_3 , and y_4 . The metric of the factor is set automatically by the program by fixing the first factor loading to one. This option can be overridden. The residual variances of the factor indicators are estimated and the residuals are not correlated as the default. The ON statement describes the linear regression of fw on the individual-level covariates x_1 and x_2 . The residual variance of the factor is estimated as the default. The intercept of the factor is fixed at zero.

In the between part of the model, the BY statement specifies that fb is measured by the random intercepts y_1 , y_2 , y_3 , and y_4 . The metric of the factor is set automatically by the program by fixing the first factor loading to one. This option can be overridden. The residual variances of the factor indicators are set to zero. The ON statement describes the regression of fb on the cluster-level covariate w . The residual variance of the factor is estimated as the default. The intercept of the factor is fixed at zero as the default. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. An explanation of the other commands can be found in Example 9.1.

EXAMPLE 9.7: TWO-LEVEL CFA WITH CATEGORICAL FACTOR INDICATORS AND COVARIATES

```

TITLE:      this is an example of a two-level CFA with
            categorical factor indicators and
            covariates
DATA:      FILE IS ex9.7.dat;
VARIABLE:  NAMES ARE u1-u4 x1 x2 w clus;
            CATEGORICAL = u1-u4;
            WITHIN = x1 x2;
            BETWEEN = w;
            CLUSTER = clus;
            MISSING = ALL (999);
ANALYSIS:  TYPE = TWOLEVEL;
MODEL:
            %WITHIN%
            fw BY u1-u4;
            fw ON x1 x2;
            %BETWEEN%
            fb BY u1-u4;
            fb ON w;
OUTPUT:    TECH1 TECH8;

```

The difference between this example and Example 9.6 is that the factor indicators are binary or ordered categorical (ordinal) variables instead of continuous variables. The CATEGORICAL option is used to specify which dependent variables are treated as binary or ordered categorical (ordinal) variables in the model and its estimation. In the example above, all four factor indicators are binary or ordered categorical. The program determines the number of categories for each indicator. The default estimator for this type of analysis is maximum likelihood with robust standard errors using a numerical integration algorithm. Note that numerical integration becomes increasingly more computationally demanding as the number of factors and the sample size increase. In this example, two dimensions of integration are used with a total of 225 integration points. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator.

In the between part of the model, the residual variances of the random intercepts of the categorical factor indicators are fixed at zero as the default because the residual variances of random intercepts are often very small and require one dimension of numerical integration each. Weighted least squares estimation of between-level residual variances

does not require numerical integration in estimating the model. An explanation of the other commands can be found in Examples 9.1 and 9.6.

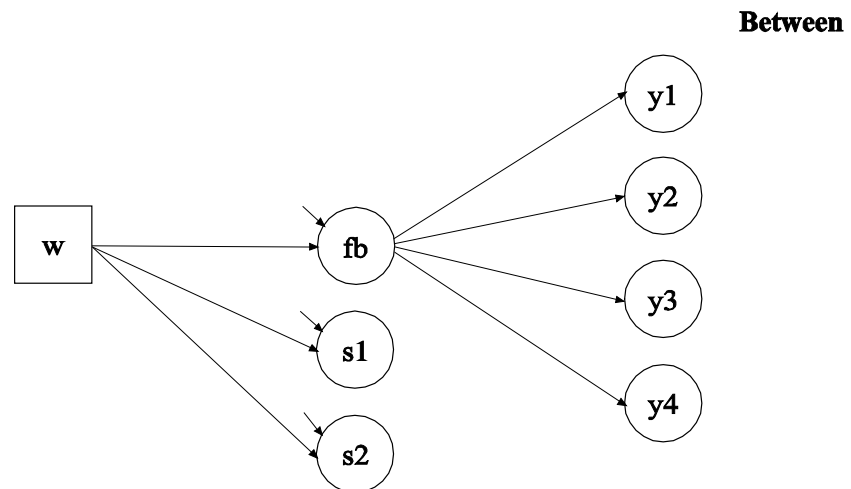
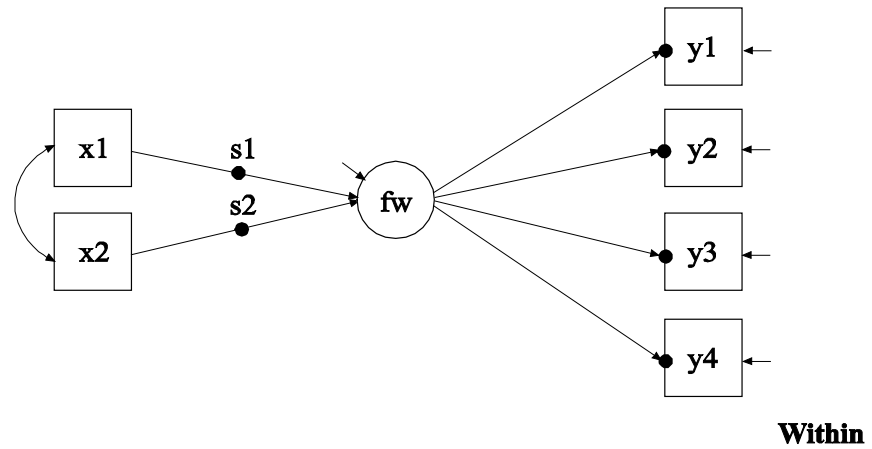
EXAMPLE 9.8: TWO-LEVEL CFA WITH CONTINUOUS FACTOR INDICATORS, COVARIATES, AND RANDOM SLOPES

```

TITLE:      this is an example of a two-level CFA with
            continuous factor indicators, covariates,
            and random slopes
DATA:       FILE IS ex9.8.dat;
VARIABLE:   NAMES ARE y1-y4 x1 x2 w clus;
            CLUSTER = clus;
            WITHIN = x1 x2;
            BETWEEN = w;
ANALYSIS:   TYPE = TWOLEVEL RANDOM;
MODEL:
            %WITHIN%
            fw BY y1-y4;
            s1 | fw ON x1;
            s2 | fw ON x2;
            %BETWEEN%
            fb BY y1-y4;
            y1-y4@0;
            fb s1 s2 ON w;

```

Examples: Multilevel Modeling With Complex Survey Data



The difference between this example and Example 9.6 is that the model has random slopes in addition to random intercepts and the random slopes are regressed on a cluster-level covariate. In the within part of the model, the filled circles on the arrows from x_1 and x_2 to fw represent random slopes that are referred to as s_1 and s_2 in the between part of the model. In the between part of the model, the random slopes are shown in circles because they are latent variables that vary across clusters.

CHAPTER 9

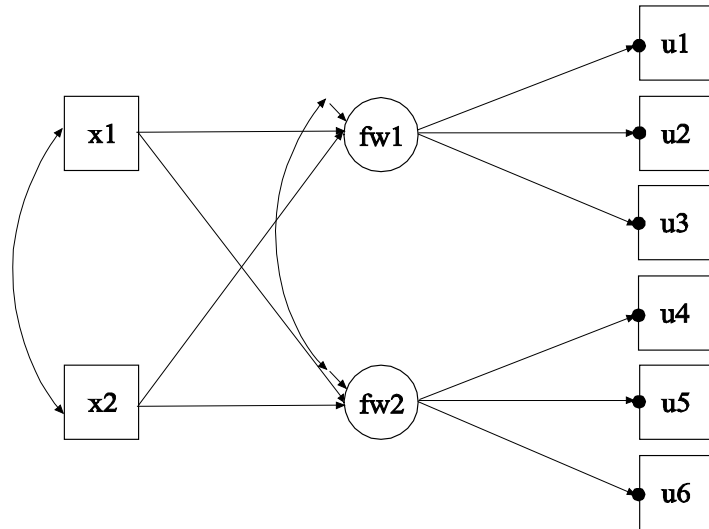
In the within part of the model, the | symbol is used in conjunction with TYPE=RANDOM to name and define the random slope variables in the model. The name on the left-hand side of the | symbol names the random slope variable. The statement on the right-hand side of the | symbol defines the random slope variable. Random slopes are defined using the ON option. In the first | statement, the random slope s1 is defined by the linear regression of the factor fw on the individual-level covariate x1. In the second | statement, the random slope s2 is defined by the linear regression of the factor fw on the individual-level covariate x2. The within-level residual variance of f1 is estimated as the default.

In the between part of the model, the ON statement describes the linear regressions of fb, s1, and s2 on the cluster-level covariate w. The residual variances of fb, s1, and s2 are estimated as the default. The residuals are not correlated as the default. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. An explanation of the other commands can be found in Examples 9.1 and 9.6.

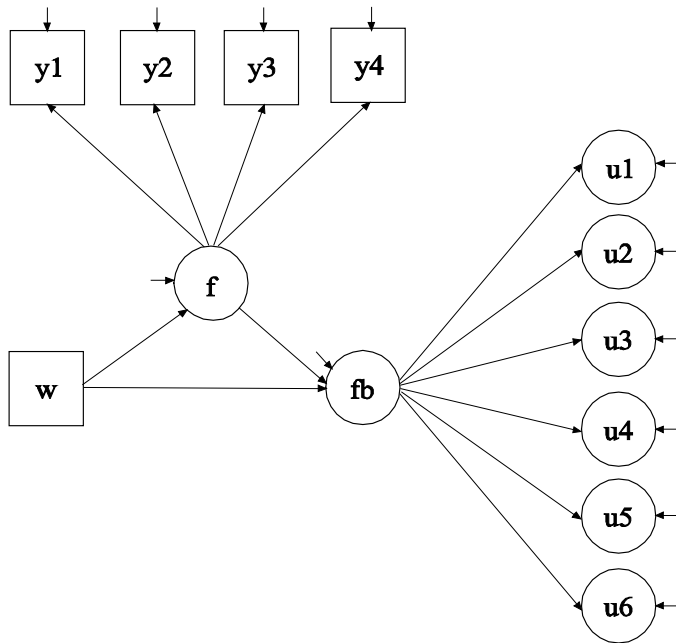
**EXAMPLE 9.9: TWO-LEVEL SEM WITH CATEGORICAL
FACTOR INDICATORS ON THE WITHIN LEVEL AND
CLUSTER-LEVEL CONTINUOUS OBSERVED AND RANDOM
INTERCEPT FACTOR INDICATORS ON THE BETWEEN
LEVEL**

```
TITLE:      this is an example of a two-level SEM with
            categorical factor indicators on the
            within level and cluster-level continuous
            observed and random intercept factor
            indicators on the between level
DATA:      FILE IS ex9.9.dat;
VARIABLE:  NAMES ARE u1-u6 y1-y4 x1 x2 w clus;
            CATEGORICAL = u1-u6;
            WITHIN = x1 x2;
            BETWEEN = w y1-y4;
            CLUSTER IS clus;
ANALYSIS:  TYPE IS TWOLEVEL;
            ESTIMATOR = WLSMV;
MODEL:
            %WITHIN%
            fw1 BY u1-u3;
            fw2 BY u4-u6;
            fw1 fw2 ON x1 x2;
            %BETWEEN%
            fb BY u1-u6;
            f BY y1-y4;
            fb ON w f;
            f ON w;
SAVEDATA:  SWMATRIX = ex9.9sw.dat;
```

CHAPTER 9



Within



Between

Examples: Multilevel Modeling With Complex Survey Data

In this example, the model with two within factors and two between factors shown in the picture above is estimated. The within-level factor indicators are categorical. In the within part of the model, the filled circles at the end of the arrows from the within factor `fw1` to `u1`, `u2`, and `u3` and `fw2` to `u4`, `u5`, and `u6` represent random intercepts that are referred to as `u1`, `u2`, `u3`, `u4`, `u5`, and `u6` in the between part of the model. In the between part of the model, the random intercepts are shown in circles because they are continuous latent variables that vary across clusters. The random intercepts are indicators of the between factor `fb`. This example illustrates the common finding of fewer between factors than within factors for the same set of factor indicators. The between factor `f` has observed cluster-level continuous variables as factor indicators.

By specifying `ESTIMATOR=WLSMV`, a robust weighted least squares estimator using a diagonal weight matrix will be used. The default estimator for this type of analysis is maximum likelihood with robust standard errors using a numerical integration algorithm. Note that numerical integration becomes increasingly more computationally demanding as the number of factors and the sample size increase. In this example, three dimensions of integration would be used with a total of 3,375 integration points. For models with many dimensions of integration and categorical outcomes, the weighted least squares estimator may improve computational speed. The `ESTIMATOR` option of the `ANALYSIS` command can be used to select a different estimator.

In the within part of the model, the first `BY` statement specifies that `fw1` is measured by `u1`, `u2`, and `u3`. The second `BY` statement specifies that `fw2` is measured by `u4`, `u5`, and `u6`. The metric of the factors are set automatically by the program by fixing the first factor loading for each factor to one. This option can be overridden. Residual variances of the latent response variables of the categorical factor indicators are not parameters in the model. They are fixed at one in line with the Theta parameterization. Residuals are not correlated as the default. The `ON` statement describes the linear regressions of `fw1` and `fw2` on the individual-level covariates `x1` and `x2`. The residual variances of the factors are estimated as the default. The residuals of the factors are correlated as the default because residuals are correlated for latent variables that do not influence any other variable in the model except their own indicators. The intercepts of the factors are fixed at zero as the default.

CHAPTER 9

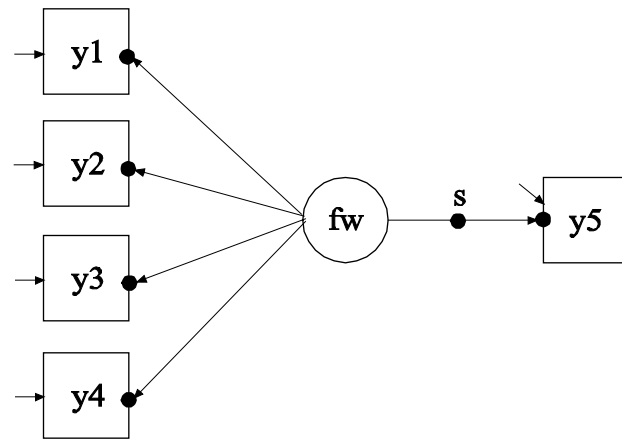
In the between part of the model, the first BY statement specifies that *fb* is measured by the random intercepts *u1*, *u2*, *u3*, *u4*, *u5*, and *u6*. The metric of the factor is set automatically by the program by fixing the first factor loading to one. This option can be overridden. The residual variances of the factor indicators are estimated and the residuals are not correlated as the default. Unlike maximum likelihood estimation, weighted least squares estimation of between-level residual variances does not require numerical integration in estimating the model. The second BY statement specifies that *f* is measured by the cluster-level factor indicators *y1*, *y2*, *y3*, and *y4*. The residual variances of the factor indicators are estimated and the residuals are not correlated as the default. The first ON statement describes the linear regression of *fb* on the cluster-level covariate *w* and the factor *f*. The second ON statement describes the linear regression of *f* on the cluster-level covariate *w*. The residual variances of the factors are estimated as the default. The intercepts of the factors are fixed at zero as the default.

The `SWMATRIX` option of the `SAVEDATA` command is used with `TYPE=TWOLEVEL` and weighted least squares estimation to specify the name and location of the file that contains the within- and between-level sample statistics and their corresponding estimated asymptotic covariance matrix. It is recommended to save this information and use it in subsequent analyses along with the raw data to reduce computational time during model estimation. An explanation of the other commands can be found in Example 9.1.

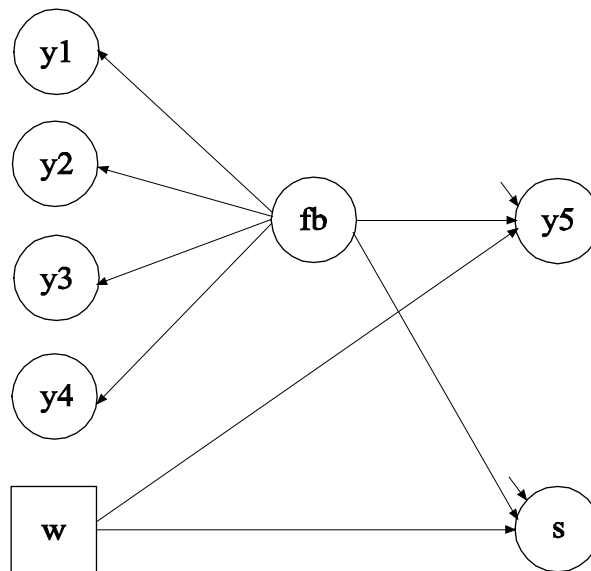
EXAMPLE 9.10: TWO-LEVEL SEM WITH CONTINUOUS FACTOR INDICATORS AND A RANDOM SLOPE FOR A FACTOR

```
TITLE:      this is an example of a two-level SEM with
             continuous factor indicators and a random
             slope for a factor
DATA:       FILE IS ex9.10.dat;
VARIABLE:   NAMES ARE y1-y5 w clus;
             BETWEEN = w;
             CLUSTER = clus;
ANALYSIS:   TYPE = TWOLEVEL RANDOM;
             ALGORITHM = INTEGRATION;
             INTEGRATION = 10;
MODEL:
             %WITHIN%
             fw BY y1-y4;
             s | y5 ON fw;
             %BETWEEN%
             fb BY y1-y4;
             y1-y4@0;
             y5 s ON fb w;
OUTPUT:     TECH1 TECH8;
```

CHAPTER 9



Within



Between

In this example, the two-level SEM with continuous factor indicators shown in the picture above is estimated. In the within part of the model, the filled circles at the end of the arrows from fw to the factor indicators y1, y2, y3, and y4 and the filled circle at the end of the arrow from fw to y5 represent random intercepts that are referred to as y1, y2, y3, y4, and y5 in the between part of the model. The filled circle on the arrow from fw to y5 represents a random slope that is referred to as s in the between

Examples: Multilevel Modeling With Complex Survey Data

part of the model. In the between part of the model, the random intercepts and random slope are shown in circles because they are continuous latent variables that vary across clusters.

By specifying `TYPE=TWOLEVEL RANDOM` in the `ANALYSIS` command, a multilevel model with random intercepts and random slopes will be estimated. By specifying `ALGORITHM=INTEGRATION`, a maximum likelihood estimator with robust standard errors using a numerical integration algorithm will be used. Note that numerical integration becomes increasingly more computationally demanding as the number of factors and the sample size increase. In this example, four dimensions of integration are used with a total of 10,000 integration points. The `INTEGRATION` option of the `ANALYSIS` command is used to change the number of integration points per dimension from the default of 15 to 10. The `ESTIMATOR` option of the `ANALYSIS` command can be used to select a different estimator.

In the within part of the model, the `BY` statement specifies that `fw` is measured by the factor indicators `y1`, `y2`, `y3`, and `y4`. The metric of the factor is set automatically by the program by fixing the first factor loading in each `BY` statement to one. This option can be overridden. The residual variances of the factor indicators are estimated and the residuals are uncorrelated as the default. The variance of the factor is estimated as the default.

In the within part of the model, the `|` symbol is used in conjunction with `TYPE=RANDOM` to name and define the random slope variables in the model. The name on the left-hand side of the `|` symbol names the random slope variable. The statement on the right-hand side of the `|` symbol defines the random slope variable. Random slopes are defined using the `ON` option. In the `|` statement, the random slope `s` is defined by the linear regression of the dependent variable `y5` on the within factor `fw`. The within-level residual variance of `y5` is estimated as the default.

In the between part of the model, the `BY` statement specifies that `fb` is measured by the random intercepts `y1`, `y2`, `y3`, and `y4`. The metric of the factor is set automatically by the program by fixing the first factor loading in the `BY` statement to one. This option can be overridden. The residual variances of the factor indicators are fixed at zero. The variance of the factor is estimated as the default. The `ON` statement describes the linear regressions of the random intercept `y5` and the random slope `s` on

the factor fb and the cluster-level covariate w. The intercepts and residual variances of y5 and s are estimated and their residuals are uncorrelated as the default.

The OUTPUT command is used to request additional output not included as the default. The TECH1 option is used to request the arrays containing parameter specifications and starting values for all free parameters in the model. The TECH8 option is used to request that the optimization history in estimating the model be printed in the output. TECH8 is printed to the screen during the computations as the default. TECH8 screen printing is useful for determining how long the analysis takes. An explanation of the other commands can be found in Example 9.1.

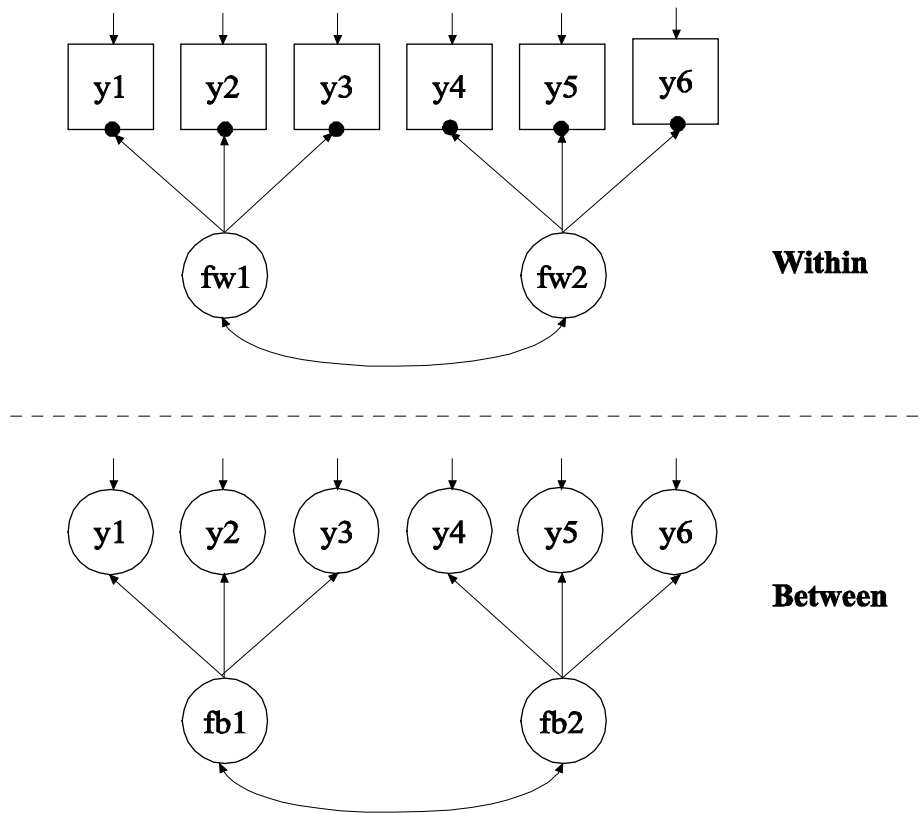
EXAMPLE 9.11: TWO-LEVEL MULTIPLE GROUP CFA WITH CONTINUOUS FACTOR INDICATORS

```

TITLE:      this is an example of a two-level
             multiple group CFA with continuous
             factor indicators
DATA:       FILE IS ex9.11.dat;
VARIABLE:   NAMES ARE y1-y6 g clus;
             GROUPING = g (1 = g1 2 = g2);
             CLUSTER = clus;
ANALYSIS:   TYPE = TWOLEVEL;
MODEL:
             %WITHIN%
             fw1 BY y1-y3;
             fw2 BY y4-y6;
             %BETWEEN%
             fb1 BY y1-y3;
             fb2 BY y4-y6;
MODEL g2:   %WITHIN%
             fw1 BY y2-y3;
             fw2 BY y5-y6;

```

Examples: Multilevel Modeling With Complex Survey Data



In this example, the two-level multiple group CFA with continuous factor indicators shown in the picture above is estimated. In the within part of the model, the filled circles at the end of the arrows from the within factors fw1 to y1, y2, and y3 and fw2 to y4, y5, and y6 represent random intercepts that are referred to as y1, y2, y3, y4, y5, and y6 in the between part of the model. In the between part of the model, the random intercepts are shown in circles because they are continuous latent variables that vary across clusters. The random intercepts are indicators of the between factors fb1 and fb2.

The GROUPING option of the VARIABLE command is used to identify the variable in the data set that contains information on group membership when the data for all groups are stored in a single data set. The information in parentheses after the grouping variable name assigns labels to the values of the grouping variable found in the data set. In the example above, observations with g equal to 1 are assigned the label g1,

CHAPTER 9

and individuals with g equal to 2 are assigned the label $g2$. These labels are used in conjunction with the MODEL command to specify model statements specific to each group. The grouping variable should be a cluster-level variable.

In multiple group analysis, two variations of the MODEL command are used. They are MODEL and MODEL followed by a label. MODEL describes the model to be estimated for all groups. The factor loadings and intercepts are held equal across groups as the default to specify measurement invariance. MODEL followed by a label describes differences between the overall model and the model for the group designated by the label.

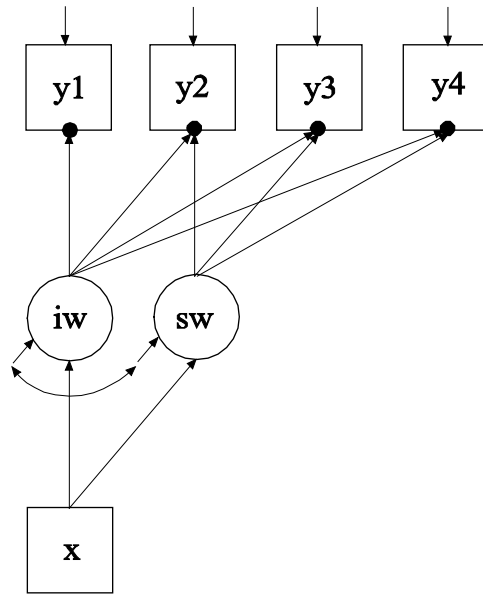
In the within part of the model, the BY statements specify that $fw1$ is measured by $y1$, $y2$, and $y3$, and $fw2$ is measured by $y4$, $y5$, and $y6$. The metric of the factors is set automatically by the program by fixing the first factor loading in each BY statement to one. This option can be overridden. The variances of the factors are estimated as the default. The factors $fw1$ and $fw2$ are correlated as the default because they are independent (exogenous) variables. In the between part of the model, the BY statements specify that $fb1$ is measured by $y1$, $y2$, and $y3$, and $fb2$ is measured by $y4$, $y5$, and $y6$. The metric of the factor is set automatically by the program by fixing the first factor loading in each BY statement to one. This option can be overridden. The variances of the factors are estimated as the default. The factors $fb1$ and $fb2$ are correlated as the default because they are independent (exogenous) variables.

In the group-specific MODEL command for group 2, by specifying the within factor loadings for $fw1$ and $fw2$, the default equality constraints are relaxed and the factor loadings are no longer held equal across groups. The factor indicators that are fixed at one remain the same, in this case $y1$ and $y4$. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. An explanation of the other commands can be found in Example 9.1.

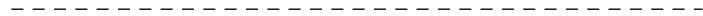
EXAMPLE 9.12: TWO-LEVEL GROWTH MODEL FOR A CONTINUOUS OUTCOME (THREE-LEVEL ANALYSIS)

```
TITLE:      this is an example of a two-level growth
            model for a continuous outcome (three-
            level analysis)
DATA:      FILE IS ex9.12.dat;
VARIABLE:  NAMES ARE y1-y4 x w clus;
            WITHIN = x;
            BETWEEN = w;
            CLUSTER = clus;
ANALYSIS:  TYPE = TWOLEVEL;
MODEL:
            %WITHIN%
            iw sw | y1@0 y2@1 y3@2 y4@3;
            y1-y4 (1);
            iw sw ON x;
            %BETWEEN%
            ib sb | y1@0 y2@1 y3@2 y4@3;
            y1-y4@0;
            ib sb ON w;
```

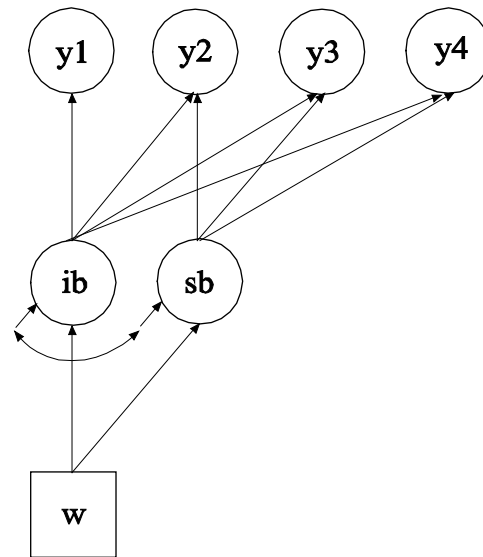
CHAPTER 9



Within



Between



Examples: Multilevel Modeling With Complex Survey Data

In this example, the two-level growth model for a continuous outcome (three-level analysis) shown in the picture above is estimated. In the within part of the model, the filled circles at the end of the arrows from the within growth factors i_w and s_w to y_1 , y_2 , y_3 , and y_4 represent random intercepts that are referred to as y_1 , y_2 , y_3 , and y_4 in the between part of the model. In the between part of the model, the random intercepts are shown in circles because they are continuous latent variables that vary across clusters.

In the within part of the model, the `|` statement names and defines the within intercept and slope factors for the growth model. The names i_w and s_w on the left-hand side of the `|` symbol are the names of the intercept and slope growth factors, respectively. The values on the right-hand side of the `|` symbol are the time scores for the slope growth factor. The time scores of the slope growth factor are fixed at 0, 1, 2, and 3 to define a linear growth model with equidistant time points. The zero time score for the slope growth factor at time point one defines the intercept growth factor as an initial status factor. The coefficients of the intercept growth factor are fixed at one as part of the growth model parameterization. The residual variances of the outcome variables are constrained to be equal over time in line with conventional multilevel growth modeling. This is done by placing (1) after them. The residual covariances of the outcome variables are fixed at zero as the default. Both of these restrictions can be overridden. The `ON` statement describes the linear regressions of the growth factors on the individual-level covariate x . The residual variances of the growth factors are free to be estimated as the default. The residuals of the growth factors are correlated as the default because residuals are correlated for latent variables that do not influence any other variable in the model except their own indicators.

In the between part of the model, the `|` statement names and defines the between intercept and slope factors for the growth model. The names i_b and s_b on the left-hand side of the `|` symbol are the names of the intercept and slope growth factors, respectively. The values on the right-hand side of the `|` symbol are the time scores for the slope growth factor. The time scores of the slope growth factor are fixed at 0, 1, 2, and 3 to define a linear growth model with equidistant time points. The zero time score for the slope growth factor at time point one defines the intercept factor as an initial status factor. The coefficients of the intercept growth factor are fixed at one as part of the growth model parameterization. The

residual variances of the outcome variables are fixed at zero on the between level in line with conventional multilevel growth modeling. These residual variances can be estimated. The ON statement describes the linear regressions of the growth factors on the cluster-level covariate *w*. The residual variances and the residual covariance of the growth factors are free to be estimated as the default.

In the parameterization of the growth model shown here, the intercepts of the outcome variable at the four time points are fixed at zero as the default. The intercepts of the growth factors are estimated as the default in the between part of the model. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. An explanation of the other commands can be found in Example 9.1.

EXAMPLE 9.13: TWO-LEVEL GROWTH MODEL FOR A CATEGORICAL OUTCOME (THREE-LEVEL ANALYSIS)

```

TITLE:      this is an example of a two-level
            growth model for a categorical outcome
            (three-level analysis)
DATA:      FILE IS ex9.13.dat;
VARIABLE:  NAMES ARE u1-u4 x w clus;
            CATEGORICAL = u1-u4;
            WITHIN = x;
            BETWEEN = w;
            CLUSTER = clus;
ANALYSIS:  TYPE = TWOLEVEL;
            INTEGRATION = 7;
MODEL:
            %WITHIN%
            iw sw | u1@0 u2@1 u3@2 u4@3;
            iw sw ON x;
            %BETWEEN%
            ib sb | u1@0 u2@1 u3@2 u4@3;
            ib sb ON w;
OUTPUT:    TECH1 TECH8;

```

The difference between this example and Example 9.12 is that the outcome variable is a binary or ordered categorical (ordinal) variable instead of a continuous variable.

Examples: Multilevel Modeling With Complex Survey Data

The CATEGORICAL option is used to specify which dependent variables are treated as binary or ordered categorical (ordinal) variables in the model and its estimation. In the example above, u1, u2, u3, and u4 are binary or ordered categorical variables. They represent the outcome measured at four equidistant occasions.

The default estimator for this type of analysis is maximum likelihood with robust standard errors using a numerical integration algorithm. Note that numerical integration becomes increasingly more computationally demanding as the number of factors and the sample size increase. In this example, four dimensions of integration are used with a total of 2,401 integration points. The INTEGRATION option of the ANALYSIS command is used to change the number of integration points per dimension from the default of 15 to 7. The ESTIMATOR option of the ANALYSIS command can be used to select a different estimator. For models with many dimensions of integration and categorical outcomes, the weighted least squares estimator may improve computational speed.

In the parameterization of the growth model shown here, the thresholds of the outcome variable at the four time points are held equal as the default and are estimated in the between part of the model. The intercept of the intercept growth factor is fixed at zero. The intercept of the slope growth factor is estimated as the default in the between part of the model. The residual variances of the growth factors are estimated as the default. The residuals of the growth factors are correlated as the default because residuals are correlated for latent variables that do not influence any other variable in the model except their own indicators. On the between level, the residual variances of the random intercepts u1, u2, u3, and u4 are fixed at zero as the default.

The OUTPUT command is used to request additional output not included as the default. The TECH1 option is used to request the arrays containing parameter specifications and starting values for all free parameters in the model. The TECH8 option is used to request that the optimization history in estimating the model be printed in the output. TECH8 is printed to the screen during the computations as the default. TECH8 screen printing is useful for determining how long the analysis takes. An explanation of the other commands can be found in Examples 9.1 and 9.12.

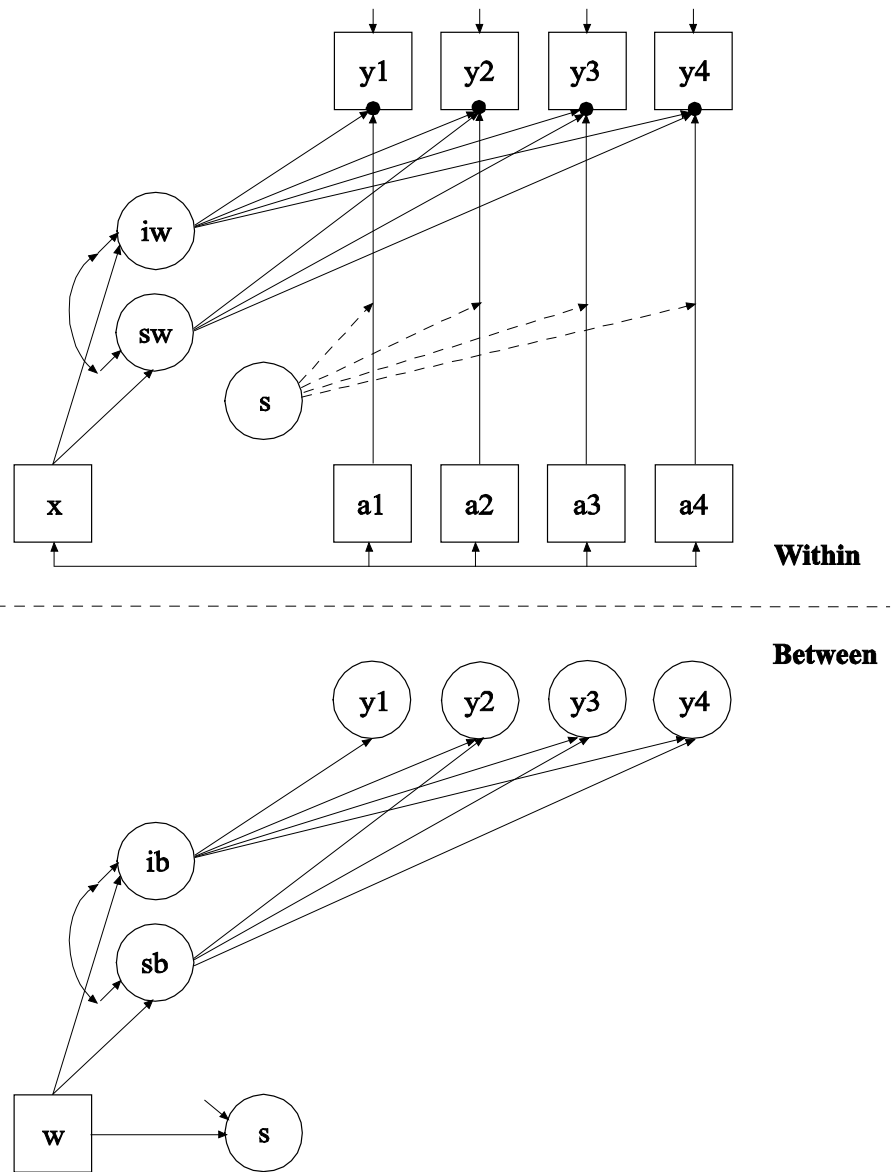
EXAMPLE 9.14: TWO-LEVEL GROWTH MODEL FOR A CONTINUOUS OUTCOME (THREE-LEVEL ANALYSIS) WITH VARIATION ON BOTH THE WITHIN AND BETWEEN LEVELS FOR A RANDOM SLOPE OF A TIME-VARYING COVARIATE

```

TITLE:      this is an example of a two-level growth
            model for a continuous outcome (three-
            level analysis) with variation on both the
            within and between levels for a random
            slope of a time-varying covariate
DATA:      FILE IS ex9.14.dat;
VARIABLE:  NAMES ARE y1-y4 x a1-a4 w clus;
            WITHIN = x a1-a4;
            BETWEEN = w;
            CLUSTER = clus;
ANALYSIS:  TYPE = TWOLEVEL RANDOM;
            ALGORITHM = INTEGRATION;
            INTEGRATION = 10;
MODEL:
            %WITHIN%
            iw sw | y1@0 y2@1 y3@2 y4@3;
            y1-y4 (1);
            iw sw ON x;
            s* | y1 ON a1;
            s* | y2 ON a2;
            s* | y3 ON a3;
            s* | y4 ON a4;
            %BETWEEN%
            ib sb | y1@0 y2@1 y3@2 y4@3;
            y1-y4@0;
            ib sb s ON w;
OUTPUT:    TECH1 TECH8;

```

Examples: Multilevel Modeling With Complex Survey Data



The difference between this example and Example 9.12 is that the model includes an individual-level time-varying covariate with a random slope that varies on both the within and between levels. In the within part of the model, the filled circles at the end of the arrows from a1 to y1, a2 to y2, a3 to y3, and a4 to y4 represent random intercepts that are referred to

CHAPTER 9

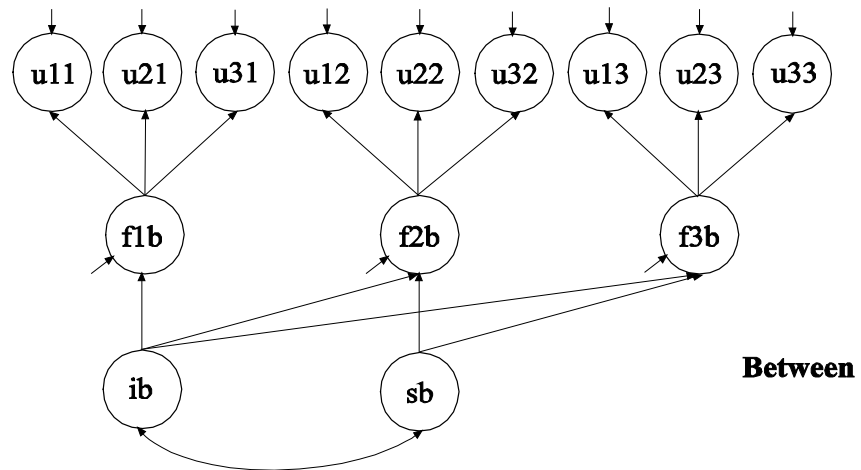
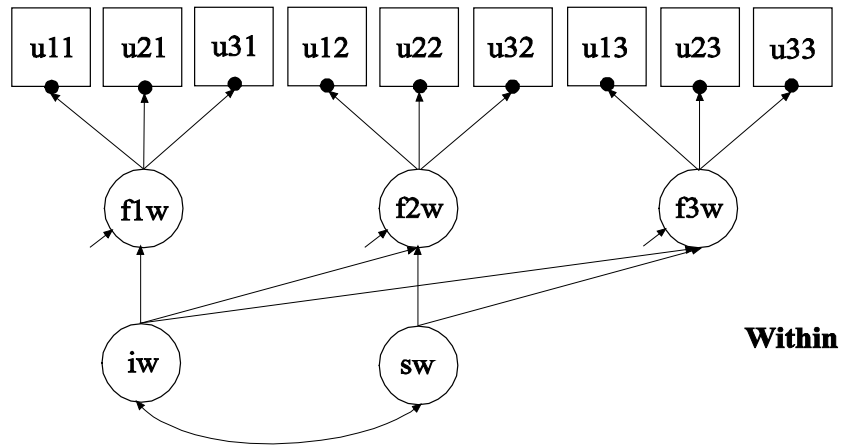
as y_1 , y_2 , y_3 , and y_4 in the between part of the model. In the between part of the model, the random intercepts are shown in circles because they are continuous latent variables that vary across classes. The broken arrows from s to the arrows from a_1 to y_1 , a_2 to y_2 , a_3 to y_3 , and a_4 to y_4 indicate that the slopes in these regressions are random. The s is shown in a circle in both the within and between parts of the model to represent a decomposition of the random slope into its within and between components.

By specifying `TYPE=TWOLEVEL RANDOM` in the `ANALYSIS` command, a multilevel model with random intercepts and random slopes will be estimated. By specifying `ALGORITHM=INTEGRATION`, a maximum likelihood estimator with robust standard errors using a numerical integration algorithm will be used. Note that numerical integration becomes increasingly more computationally demanding as the number of factors and the sample size increase. In this example, four dimensions of integration are used with a total of 10,000 integration points. The `INTEGRATION` option of the `ANALYSIS` command is used to change the number of integration points per dimension from the default of 15 to 10. The `ESTIMATOR` option of the `ANALYSIS` command can be used to select a different estimator.

The `|` symbol is used in conjunction with `TYPE=RANDOM` to name and define the random slope variables in the model. The name on the left-hand side of the `|` symbol names the random slope variable. The statement on the right-hand side of the `|` symbol defines the random slope variable. The random slope s is defined by the linear regressions of y_1 on a_1 , y_2 on a_2 , y_3 on a_3 , and y_4 on a_4 . Random slopes with the same name are treated as one variable during model estimation. The random intercepts for these regressions are referred to by using the name of the dependent variables in the regressions, that is, y_1 , y_2 , y_3 , and y_4 . The asterisk (*) following the s specifies that s will have variation on both the within and between levels. Without the asterisk (*), s would have variation on only the between level. An explanation of the other commands can be found in Examples 9.1 and 9.12.

EXAMPLE 9.15: TWO-LEVEL MULTIPLE INDICATOR GROWTH MODEL WITH CATEGORICAL OUTCOMES (THREE-LEVEL ANALYSIS)

```
TITLE:      this is an example of a two-level multiple
            indicator growth model with categorical
            outcomes (three-level analysis)
DATA:      FILE IS ex9.15.dat;
VARIABLE:  NAMES ARE u11 u21 u31 u12 u22 u32 u13 u23
            u33 clus;
            CATEGORICAL = u11-u33;
            CLUSTER = clus;
ANALYSIS:  TYPE IS TWOLEVEL;
            ESTIMATOR = WLSM;
MODEL:
            %WITHIN%
            f1w BY u11
            u21-u31 (1-2);
            f2w BY u12
            u22-u32 (1-2);
            f3w BY u13
            u23-u33 (1-2);
            iw sw | f1w@0 f2w@1 f3w@2;
            %BETWEEN%
            f1b BY u11
            u21-u31 (1-2);
            f2b BY u12
            u22-u32 (1-2);
            f3b BY u13
            u23-u33 (1-2);
            [u11$1 u12$1 u13$1] (3);
            [u21$1 u22$1 u23$1] (4);
            [u31$1 u32$1 u33$1] (5);
            ib sb | f1b@0 f2b@1 f3b@2;
            [f1b-f3b@0 ib@0 sb];
            f1b-f3b (6);
SAVEDATA:  SWMATRIX = ex9.15sw.dat;
```



In this example, the two-level multiple indicator growth model with categorical outcomes (three-level analysis) shown in the picture above is estimated. The picture shows a factor measured by three indicators at three time points. In the within part of the model, the filled circles at the end of the arrows from the within factors f_{1w} to $u_{11}, u_{21},$ and u_{31} ; f_{2w} to $u_{12}, u_{22},$ and u_{32} ; and f_{3w} to $u_{13}, u_{23},$ and u_{33} represent random intercepts that are referred to as $u_{11}, u_{21}, u_{31}, u_{12}, u_{22}, u_{32}, u_{13}, u_{23},$ and u_{33} in the between part of the model. In the between part of the model, the random intercepts are continuous latent variables that vary across clusters. The random intercepts are indicators of the between factors $f_{1b}, f_{2b},$ and f_{3b} . In this model, the residual variances of the

Examples: Multilevel Modeling With Complex Survey Data

factor indicators in the between part of the model are estimated. The residuals are not correlated as the default. Taken together with the specification of equal factor loadings on the within and the between parts of the model, this implies a model where the regressions of the within factors on the growth factors have random intercepts that vary across the clusters.

By specifying `ESTIMATOR=WLSM`, a robust weighted least squares estimator using a diagonal weight matrix will be used. The default estimator for this type of analysis is maximum likelihood with robust standard errors using a numerical integration algorithm. Note that numerical integration becomes increasingly more computationally demanding as the number of factors and the sample size increase. For models with many dimensions of integration and categorical outcomes, the weighted least squares estimator may improve computational speed.

In the within part of the model, the three `BY` statements define a within-level factor at three time points. The metric of the three factors is set automatically by the program by fixing the first factor loading to one. This option can be overridden. The (1-2) following the factor loadings uses the list function to assign equality labels to these parameters. The label 1 is assigned to the factor loadings of `u21`, `u22`, and `u23` which holds these factor loadings equal across time. The label 2 is assigned to the factor loadings of `u31`, `u32`, and `u33` which holds these factor loadings equal across time. Residual variances of the latent response variables of the categorical factor indicators are not free parameters to be estimated in the model. They are fixed at one in line with the Theta parameterization. Residuals are not correlated as the default. The `|` statement names and defines the within intercept and slope growth factors for the growth model. The names `iw` and `sw` on the left-hand side of the `|` symbol are the names of the intercept and slope growth factors, respectively. The names and values on the right-hand side of the `|` symbol are the outcome and time scores for the slope growth factor. The time scores of the slope growth factor are fixed at 0, 1, and 2 to define a linear growth model with equidistant time points. The zero time score for the slope growth factor at time point one defines the intercept growth factor as an initial status factor. The coefficients of the intercept growth factor are fixed at one as part of the growth model parameterization. The variances of the growth factors are free to be estimated as the default. The covariance between the growth factors is free to be estimated as the default. The intercepts of the factors defined using `BY`

CHAPTER 9

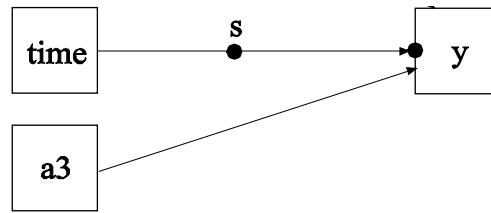
statements are fixed at zero. The residual variances of the factors are free and not held equal across time. The residuals of the factors are uncorrelated in line with the default of residuals for first-order factors.

In the between part of the model, the first three BY statements define a between-level factor at three time points. The (1-2) following the factor loadings uses the list function to assign equality labels to these parameters. The label 1 is assigned to the factor loadings of u21, u22, and u23 which holds these factor loadings equal across time as well as across levels. The label 2 is assigned to the factor loadings of u31, u32, and u33 which holds these factor loadings equal across time as well as across levels. Time-invariant thresholds for the three indicators are specified using (3), (4), and (5) following the bracket statements. The residual variances of the factor indicators are free to be estimated. The | statement names and defines the between intercept and slope growth factors for the growth model. The names ib and sb on the left-hand side of the | symbol are the names of the intercept and slope growth factors, respectively. The values on the right-hand side of the | symbol are the time scores for the slope growth factor. The time scores of the slope growth factor are fixed at 0, 1, and 2 to define a linear growth model with equidistant time points. The zero time score for the slope growth factor at time point one defines the intercept growth factor as an initial status factor. The coefficients of the intercept growth factor are fixed at one as part of the growth model parameterization. In the parameterization of the growth model shown here, the intercept growth factor mean is fixed at zero as the default for identification purposes. The variances of the growth factors are free to be estimated as the default. The covariance between the growth factors is free to be estimated as the default. The intercepts of the factors defined using BY statements are fixed at zero. The residual variances of the factors are held equal across time. The residuals of the factors are uncorrelated in line with the default of residuals for first-order factors.

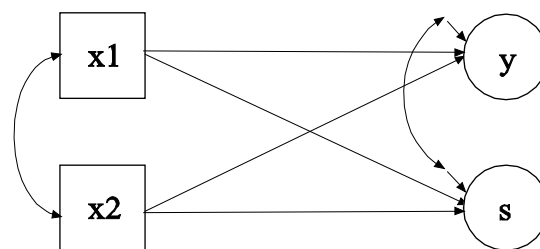
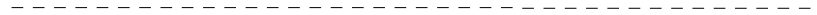
The SWMATRIX option of the SAVEDATA command is used with TYPE=TWOLEVEL and weighted least squares estimation to specify the name and location of the file that contains the within- and between-level sample statistics and their corresponding estimated asymptotic covariance matrix. It is recommended to save this information and use it in subsequent analyses along with the raw data to reduce computational time during model estimation. An explanation of the other commands can be found in Example 9.1

EXAMPLE 9.16: LINEAR GROWTH MODEL FOR A CONTINUOUS OUTCOME WITH TIME-INVARIANT AND TIME-VARYING COVARIATES CARRIED OUT AS A TWO-LEVEL GROWTH MODEL USING THE DATA WIDETOLONG COMMAND

```
TITLE:      this is an example of a linear growth
            model for a continuous outcome with time-
            invariant and time-varying covariates
            carried out as a two-level growth model
            using the DATA WIDETOLONG command
DATA:      FILE IS ex9.16.dat;
DATA WIDETOLONG:
            WIDE = y11-y14 | a31-a34;
            LONG = y | a3;
            IDVARIABLE = person;
            REPETITION = time;
VARIABLE:  NAMES ARE y11-y14 x1 x2 a31-a34;
            USEVARIABLE = x1 x2 y a3 person time;
            CLUSTER = person;
            WITHIN = time a3;
            BETWEEN = x1 x2;
ANALYSIS:  TYPE = TWOLEVEL RANDOM;
MODEL:    %WITHIN%
            s | y ON time;
            y ON a3;
            %BETWEEN%
            y s ON x1 x2;
            y WITH s;
```



Within



Between

In this example, a linear growth model for a continuous outcome with time-invariant and time-varying covariates as shown in the picture above is estimated. As part of the analysis, the DATA WIDETOLONG command is used to rearrange the data from a multivariate wide format to a univariate long format. The model is similar to the one in Example 6.10 using multivariate wide format data. The differences are that the current model restricts the within-level residual variances to be equal across time and the within-level influence of the time-varying covariate on the outcome to be equal across time.

The WIDE option of the DATA WIDETOLONG command is used to identify sets of variables in the wide format data set that are to be converted into single variables in the long format data set. These variables must be variables from the NAMES statement of the VARIABLE command. The two sets of variables y11, y12, y13, and y14 and a31, a32, a33, and a34 are identified. The LONG option is used to provide names for the new variables in the long format data set. The names y and a3 are the names of the new variables. The IDVARIABLE option is used to provide a name for the variable that provides information about the unit to which the record belongs. In univariate growth modeling, this is the person identifier which is used as a cluster variable. In this example, the name person is used. This option is not required. The

Examples: Multilevel Modeling With Complex Survey Data

default variable name is `id`. The `REPETITION` option is used to provide a name for the variable that contains information on the order in which the variables were measured. In this example, the name `time` is used. This option is not required. The default variable name is `rep`. The new variables must be mentioned on the `USEVARIABLE` statement of the `VARIABLE` command if they are used in the analysis. They must be placed after any original variables. The `USEVARIABLES` option lists the original variables `x1` and `x2` followed by the new variables `y`, `a3`, `person`, and `time`.

The `CLUSTER` option of the `VARIABLE` command is used to identify the variable that contains clustering information. In this example, the cluster variable `person` is the variable that was created using the `IDVARIABLE` option of the `DATA WIDETOLONG` command. The `WITHIN` option is used to identify the variables in the data set that are measured on the individual level and modeled only on the within level. They are specified to have no variance in the between part of the model. The `BETWEEN` option is used to identify the variables in the data set that are measured on the cluster level and modeled only on the between level. Variables not mentioned on the `WITHIN` or the `BETWEEN` statements are measured on the individual level and can be modeled on both the within and between levels.

In the within part of the model, the `|` symbol is used in conjunction with `TYPE=RANDOM` to name and define the random slope variables in the model. The name on the left-hand side of the `|` symbol names the random slope variable. The statement on the right-hand side of the `|` symbol defines the random slope variable. Random slopes are defined using the `ON` option. In the `|` statement, the random slope `s` is defined by the linear regression of the dependent variable `y` on `time`. The within-level residual variance of `y` is estimated as the default. The `ON` statement describes the linear regression of `y` on the covariate `a3`.

In the between part of the model, the `ON` statement describes the linear regressions of the random intercept `y` and the random slope `s` on the covariates `x1` and `x2`. The `WITH` statement is used to free the covariance between `y` and `s`. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The estimator option of the `ANALYSIS` command can be used to select a different estimator. An explanation of the other commands can be found in Example 9.1.

EXAMPLE 9.17: TWO-LEVEL GROWTH MODEL FOR A COUNT OUTCOME USING A ZERO-INFLATED POISSON MODEL (THREE-LEVEL ANALYSIS)

```

TITLE:      this is an example of a two-level growth
            model for a count outcome using a zero-
            inflated Poisson model (three-level
            analysis)
DATA:       FILE = ex9.17.dat;
VARIABLE:   NAMES = u1-u4 x w clus;
            COUNT = u1-u4 (i);
            CLUSTER = clus;
            WITHIN = x;
            BETWEEN = w;
ANALYSIS:   TYPE = TWOLEVEL;
            ALGORITHM = INTEGRATION;
            INTEGRATION = 10;
            MCONVERGENCE = 0.01;
MODEL:      %WITHIN%
            iw sw | u1@0 u2@1 u3@2 u4@3;
            iiw siw | u1#1@0 u2#1@1 u3#1@2 u4#1@3;
            sw@0;
            siw@0;
            iw WITH iiw;
            iw ON x;
            sw ON x;
            %BETWEEN%
            ib sb | u1@0 u2@1 u3@2 u4@3;
            iib sib | u1#1@0 u2#1@1 u3#1@2 u4#1@3;
            sb-sib@0;
            ib ON w;
OUTPUT:     TECH1 TECH8;

```

The difference between this example and Example 9.12 is that the outcome variable is a count variable instead of a continuous variable.

The COUNT option is used to specify which dependent variables are treated as count variables in the model and its estimation and whether a Poisson or zero-inflated Poisson model will be estimated. In the example above, u1, u2, u3, and u4 are count variables. The i in parentheses following u indicates that a zero-inflated Poisson model will be estimated.

Examples: Multilevel Modeling With Complex Survey Data

By specifying `ALGORITHM=INTEGRATION`, a maximum likelihood estimator with robust standard errors using a numerical integration algorithm will be used. Note that numerical integration becomes increasingly more computationally demanding as the number of factors and the sample size increase. In this example, three dimensions of integration are used with a total of 1,000 integration points. The `INTEGRATION` option of the `ANALYSIS` command is used to change the number of integration points per dimension from the default of 15 to 10. The `ESTIMATOR` option of the `ANALYSIS` command can be used to select a different estimator. The `MCONVERGENCE` option is used to change the observed-data log likelihood derivative convergence criterion for the EM algorithm from the default value of .001 to .01 because it is difficult to obtain high numerical precision in this example.

With a zero-inflated Poisson model, two growth models are estimated. In the within and between parts of the model, the first `|` statement describes the growth model for the count part of the outcome for individuals who are able to assume values of zero and above. The second `|` statement describes the growth model for the inflation part of the outcome, the probability of being unable to assume any value except zero. The binary latent inflation variable is referred to by adding to the name of the count variable the number sign (`#`) followed by the number 1. In the parameterization of the growth model for the count part of the outcome, the intercepts of the outcome variables at the four time points are fixed at zero as the default. In the parameterization of the growth model for the inflation part of the outcome, the intercepts of the outcome variable at the four time points are held equal as the default. In the within part of the model, the variances of the growth factors are estimated as the default, and the growth factor covariances are fixed at zero as the default. In the between part of the model, the mean of the growth factors for the count part of outcome are free. The mean of the intercept growth factor for the inflation part of the outcome is fixed at zero and the mean for the slope growth factor for the inflation part of the outcome is free. The variances of the growth factors are estimated as the default, and the growth factor covariances are fixed at zero as the default.

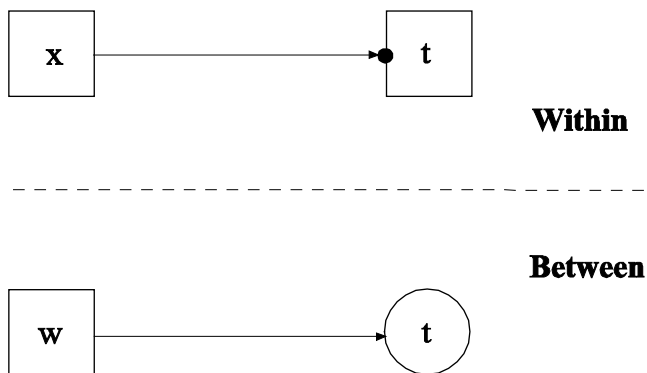
In the within part of the model, the variances of the slope growth factors `sw` and `siw` are fixed at zero. The `ON` statements describes the linear regressions of the intercept and slope growth factors `iw` and `sw` for the count part of the outcome on the covariate `x`. In the between part of the

model, the variances of the intercept growth factor iib and the slope growth factors sb and sib are fixed at zero. The ON statement describes the linear regression of the intercept growth factor ib on the covariate w. An explanation of the other commands can be found in Examples 9.1 and 9.12.

EXAMPLE 9.18: TWO-LEVEL CONTINUOUS-TIME SURVIVAL ANALYSIS USING COX REGRESSION WITH A RANDOM INTERCEPT

```

TITLE:      this is an example of a two-level
             continuous-time survival analysis using
             Cox regression with a random intercept
DATA:      FILE = ex9.18.dat;
VARIABLE:  NAMES = t x w tc clus;
             CLUSTER = clus;
             WITHIN = x;
             BETWEEN = w;
             SURVIVAL = t (ALL);
             TIMECENSORED = tc (0 = NOT 1 = RIGHT);
ANALYSIS:  TYPE = TWOLEVEL;
             BASEHAZARD = OFF;
MODEL:     %WITHIN%
             t ON x;
             %BETWEEN%
             t ON w;
    
```



Examples: Multilevel Modeling With Complex Survey Data

In this example, the two-level continuous-time survival analysis model shown in the picture above is estimated. This is the Cox regression model with a random intercept (Klein & Moeschberger, 1997; Hougaard, 2000). The profile likelihood method is used for estimation (Asparouhov et al., 2006).

The `SURVIVAL` option is used to identify the variables that contain information about time to event and to provide information about the time intervals in the baseline hazard function to be used in the analysis. The `SURVIVAL` option must be used in conjunction with the `TIMECENSORED` option. In this example, `t` is the variable that contains time to event information. By specifying the keyword `ALL` in parenthesis following the time-to-event variable, the time intervals are taken from the data. The `TIMECENSORED` option is used to identify the variables that contain information about right censoring. In this example, this variable is named `tc`. The information in parentheses specifies that the value zero represents no censoring and the value one represents right censoring. This is the default. The `BASEHAZARD` option of the `ANALYSIS` command is used with continuous-time survival analysis to specify if a non-parametric or a parametric baseline hazard function is used in the estimation of the model. The setting `OFF` specifies that a non-parametric baseline hazard function is used. This is the default.

The `MODEL` command is used to describe the model to be estimated. In multilevel models, a model is specified for both the within and between parts of the model. In the within part of the model, the loglinear regression of the time-to-event `t` on the covariate `x` is specified. In the between part of the model, the linear regression of the random intercept `t` on the cluster-level covariate `w` is specified. The default estimator for this type of analysis is maximum likelihood with robust standard errors. The estimator option of the `ANALYSIS` command can be used to select a different estimator. An explanation of the other commands can be found in Example 9.1.

CHAPTER 9