



Binomial model for the THAI data

1. Introduction to HGLM models	1
2. Description of the model	2
3. Description of the data.....	3
4. Creating the command file.....	3
5. Interpreting the output	7

1. Introduction to HGLM models

Model specification for nonlinear analysis are specified via the **Basic Settings** dialog box as shown below. Six options are currently available. For the Binomial model and Poisson (variable exposure) the TRIAL or exposure variable is selected to the right of these two options; for the multinomial and ordinal models the number of categories should also be specified to the right of these two options.

If desired, and over-dispersion option is available for binomial and Poisson models. This option is not available with Laplace estimation. To specify over-dispersion, set the σ^2 to **computed** on the **Estimation Settings** dialog box accessed via **Other Settings** on the main menu bar.

The nonlinear analysis is doubly iterative so the maximum number of macro iterations can be specified as well as the maximum number of micro iterations. Similarly, convergence criteria can be set for macro and micro iterations.

This is the first in a set of six examples illustrating HGLM models.

2. Description of the model

While the standard HLM uses a normal sampling model and an identity link function, the binary outcome model uses a binomial sampling model and a logit link. Only the level-1 models differ from the linear case.

Let Y_{ij} be the number of “successes” in m_{ij} trials. Then we write that

$$Y_{ij} | \phi_{ij} \sim B(m_{ij}, \phi_{ij}),$$

to denote that Y_{ij} has a binomial distribution with m_{ij} trials and probability of success ϕ_{ij} . According to the binomial distribution, the expected value and variance of Y_{ij} are then

$$E(Y_{ij} | \phi_{ij}) = m_{ij}\phi_{ij} \quad \text{Var}(Y_{ij} | \phi_{ij}) = m_{ij}\phi_{ij}(1 - \phi_{ij}).$$

When $m_{ij} = 1$, Y_{ij} may take on values of either zero or unity. This is a special case of the binomial distribution known as the Bernoulli distribution. HGLM allows estimation of models in which $m_{ij} = 1$ (Bernoulli case) or $m_{ij} > 1$ (other binomial cases).

When the level-1 sampling model is binomial, HGLM uses the logit link function

$$\eta_{ij} = \log\left(\frac{\phi_{ij}}{1-\phi_{ij}}\right).$$

In words, η_{ij} is the log of the odds of success. Thus if the probability of success, ϕ_{ij} , is 0.5, the odds of success is 1.0 and the log-odds or “logit” is zero. When the probability of success is less than 0.5, the odds are less than one and the logit is negative; when the probability is greater than 0.5, the odds are greater than unity and the logit is positive. Thus, while ϕ_{ij} is constrained to be in the interval (0,1), η_{ij} can take on any real value.

3. Description of the data

Data are from a national survey of primary education in Thailand (see Raudenbush & Bhumirat, 1992, for details), conducted in 1988, and yielding, for our analysis, complete data on 7516 sixth graders nested within 356 primary schools. Of interest is the probability that a child will repeat a grade during the primary years (REP1 = 1 if yes, 0 if no). It is hypothesized that the sex of the child (MALE = 1 if male, 0 if female), the child's pre-primary experience (PPED = 1 if yes, 0 if no), and the school mean SES (MSESC) will be associated with the probability of repetition. Every level-1 record corresponds to a student, with a single binary outcome per student, so the model type is Bernoulli. These data (level-1 and level-2) data files are UTHAIL1.SAV and THAI2.SAV.

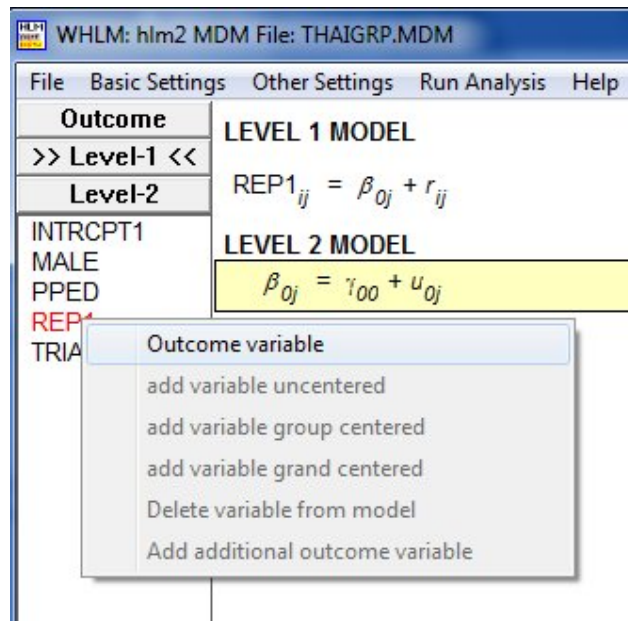
Although raw level-1 data were available, the two level-1 predictors, MALE and pre-primary experience, were categorical. For illustration of the binomial model, we reorganized these data so that each school had, potentially, four cells defined by the cross-classification of sex and pre-primary experience:

- females without pre-primary experience
- females with pre-primary experience
- males without pre-primary experience
- males with pre-primary experience

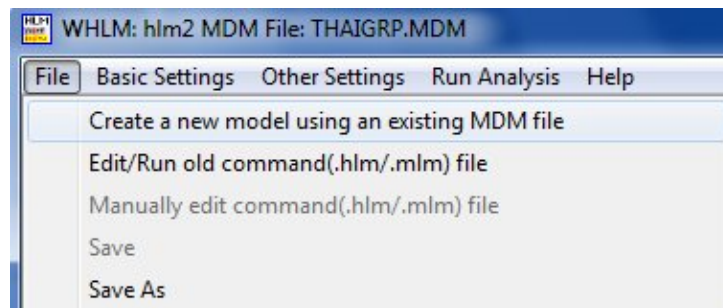
Level-1 predictors were the same as before, with MALE = 1 if male, 0 if female; PPED = 1 if pre-primary experience, 0 if not. The outcome is the number of children in a particular cell who repeated a grade, and we created a variable TRIAL, which is the number of children in each cell. In some schools there were no children of a certain type (e.g., no females with pre-primary experience). Such schools would have fewer than four cells.

4. Creating the command file

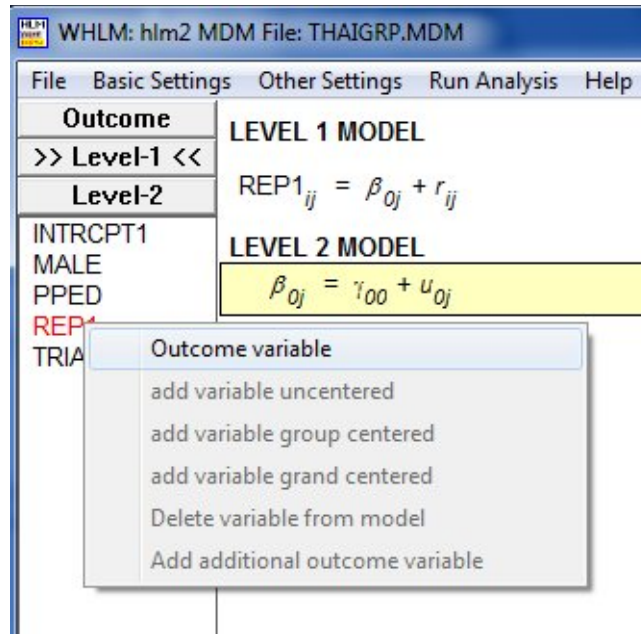
The MDM file for a HGLM model is constructed in exactly the same way as for a linear model. The procedure is described in detail for the MDM and MDM data in other examples. Using the MDM file **THAIGRP.MDM**, we set up the model as shown below.



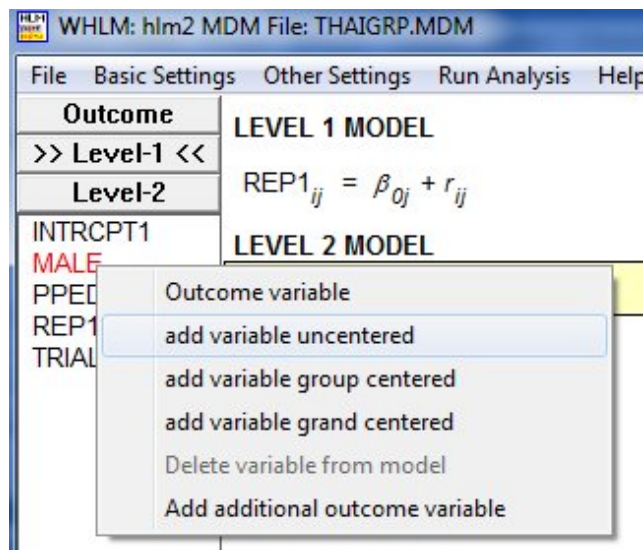
Start by selecting the **Create a new model using an existing MDM file** option from the **File** menu and open the MDM file **THAIGRP.MDM**.



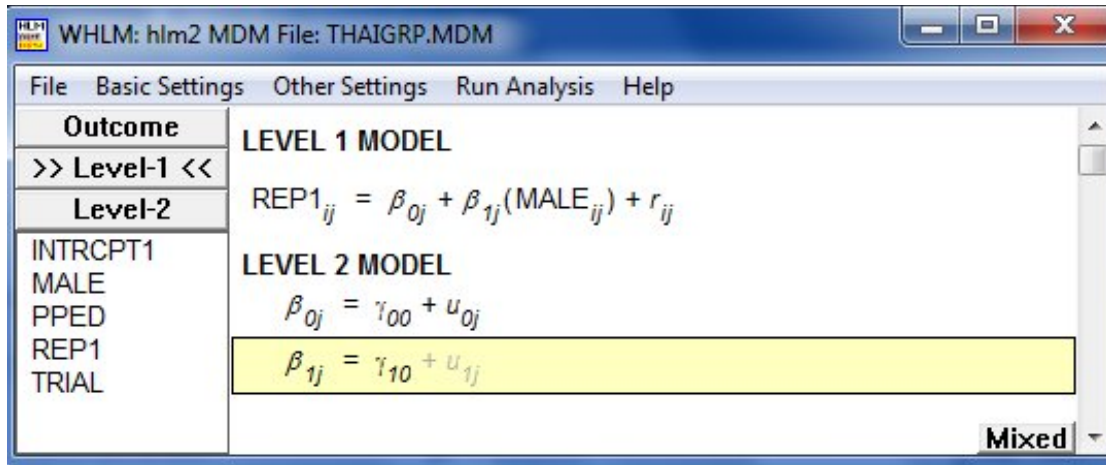
Select the outcome variable REP1 by clicking on the variable name at left and selecting **Outcome variable** from the pop-up menu.



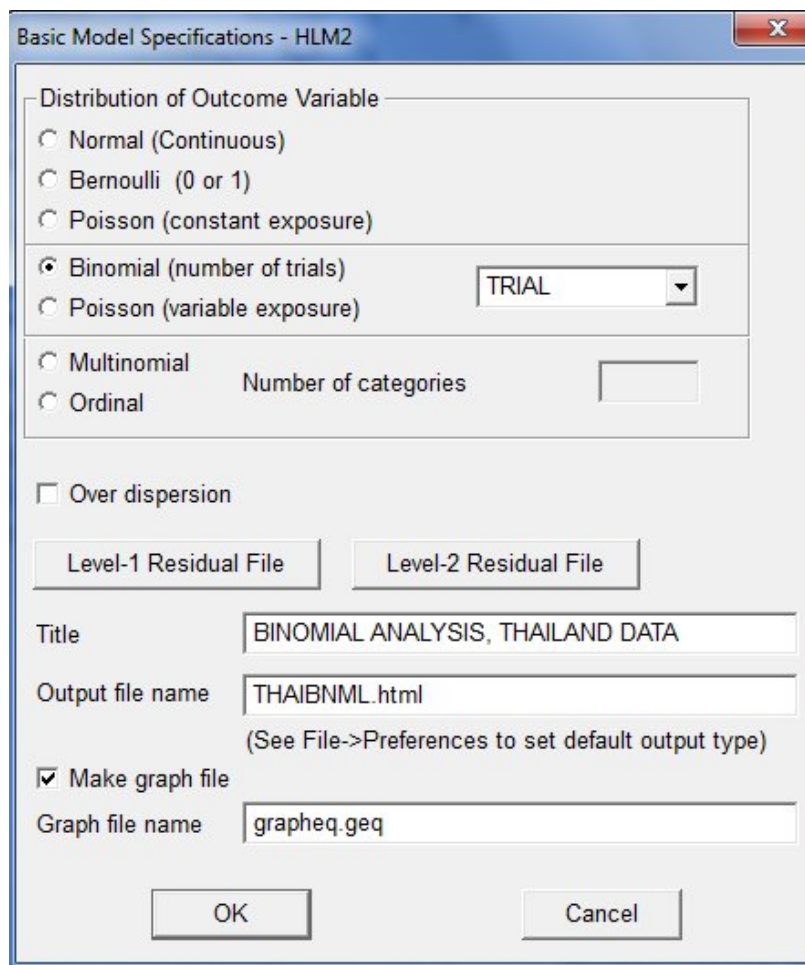
Next, add the variables MALE and PPED to the model by selecting the **add variable uncentered** option from the pop-up menu.



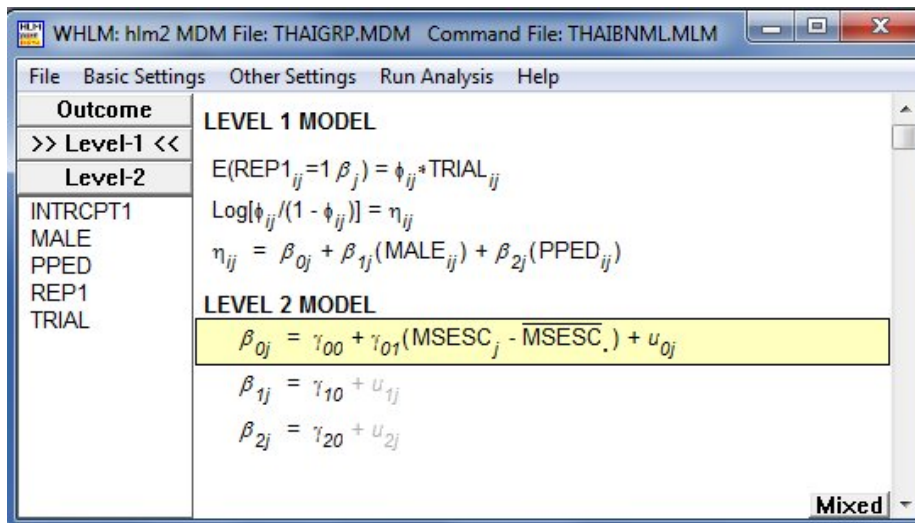
The linear model obtained is shown below. As a final step, include the level-2 predictor MDESC as **grand-mean centered** predictor on the level-2 intercept equation.



Click on **Basic Settings** on the main menu bar to indicate that the outcome has a binomial distribution. Select the variable TRIAL as indicator of the number of trials.



Click **OK** to return to the main window. The model specification is now that of a binomial model using a logit link function.



Click **OK** to return to the main window and remember to save the command file prior to running the analysis.

5. Interpreting the output

Output after convergence is shown below.

Problem Title: BINOMIAL ANALYSIS, THAILAND DATA

The data source for this run = THAIGRP.MDM

The command file for this run = C:\Users\Mathilda\AppData\Local\Temp\whlmtemp.hlm

Output file name = THAIBNML.html

The maximum number of level-1 units = 1097

The maximum number of level-2 units = 356

The maximum number of micro iterations = 50

Method of estimation: full PQL

Maximum number of macro iterations = 50

Distribution at Level-1: Binomial

The outcome variable is REP1

Summary of the model specified

Step 2 model

Level-1 Model

$$E(\text{REP1}_{ij}=1|\beta_j) = \phi_{ij} * \text{TRIAL}_{ij}$$

$$\log[\phi_{ij}/(1 - \phi_{ij})] = \eta_{ij}$$

$$\eta_{ij} = \beta_{0j} + \beta_{1j} * (\text{MALE}_{ij}) + \beta_{2j} * (\text{PPED}_{ij})$$

This is the program's way of saying that the level-1 sampling model is binomial with “TRIAL” indicating the number of trials, so that the above equation, written with subscripts and Greek letters, is

$$E(Y_{ij} | \beta_j) = m_{ij}\phi_{ij}$$

$$Var(Y_{ij} | \beta_j) = m_{ij}\phi_{ij}(1 - \phi_{ij}),$$

where m_{ij} = TRIAL.

Level-2 Model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(MSESC_j) + u_{0j}$$

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

$$\beta_{2j} = \gamma_{20}$$

MSESC has been centered around the grand mean.

$$\text{Level-1 variance} = 1/[\text{TRIAL} * \phi_{ij} (1 - \phi_{ij})]$$

In the metric of the linearized dependent variable, the level-1 variance is the reciprocal of the binomial variance,

$$m_{ij}\phi_{ij}(1 - \phi_{ij}).$$

Two sets of results are given. First of all, unit-specific results are given with and without robust standard errors. The first table provides model-based estimates of the standard errors while the second table provides robust estimates of the standard errors. Note that the two sets of standard errors are similar. If the robust and model-based standard errors are substantively different, it is recommended that the tenability of key assumptions should be investigated further.

τ
INTRCPT1,β₀ 1.28154

Standard error of τ
INTRCPT1,β₀ 0.13841

Approximate confidence intervals of tau variances
INTRCPT1 : (1.036,1.585)

Random level-1 coefficient	Reliability estimate
INTRCPT1,β ₀	0.680

The value of the log-likelihood function at iteration 2 = -1.713165E+03

Final estimation of fixed effects: (Unit-specific model)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-2.045276	0.093716	-21.824	354	<0.001
MSESC, γ_{01}	-0.254150	0.192514	-1.320	354	0.188
For MALE slope, β_1					
INTRCPT2, γ_{10}	0.508193	0.073907	6.876	739	<0.001
For PPED slope, β_2					
INTRCPT2, γ_{20}	-0.594099	0.095855	-6.198	739	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-2.045276	0.129344	(0.108,0.156)
MSESC, γ_{01}	-0.254150	0.775575	(0.531,1.133)
For MALE slope, β_1			
INTRCPT2, γ_{10}	0.508193	1.662286	(1.438,1.922)
For PPED slope, β_2			
INTRCPT2, γ_{20}	-0.594099	0.552060	(0.457,0.666)

**Final estimation of fixed effects
(Unit-specific model with robust standard errors)**

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-2.045276	0.094802	-21.574	354	<0.001
MSESC, γ_{01}	-0.254150	0.203902	-1.246	354	0.213
For MALE slope, β_1					
INTRCPT2, γ_{10}	0.508193	0.075954	6.691	739	<0.001
For PPED slope, β_2					
INTRCPT2, γ_{20}	-0.594099	0.094819	-6.266	739	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-2.045276	0.129344	(0.107,0.156)
MSESC, γ_{01}	-0.254150	0.775575	(0.519,1.158)
For MALE slope, β_1			
INTRCPT2, γ_{10}	0.508193	1.662286	(1.432,1.930)
For PPED slope, β_2			
INTRCPT2, γ_{20}	-0.594099	0.552060	(0.458,0.665)

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	d.f.	χ^2	p-value
INTRCPT1, u_0	1.13205	1.28154	354	1430.37433	<0.001

HGLM produces estimates for both the unit-specific and population-average models. “Unit-specific“ models model the expected outcome for a level-2 unit conditional on a given set of random effects. The population-average results are based on generalized least squares given the variance-covariance estimates from the unit-specific model. Moreover, HGLM produces robust standard error estimates for the population-average model (Zeger, *et al.*, 1988). These standard errors are relatively insensitive to misspecification of the variances and covariances at the two levels and to the distributional assumptions at each level. The method of estimation used in HGLM for the population-average model is equivalent to the “generalized estimating equation” (GEE) approach popularized by Zeger, *et al.* (1988).

The following differences between unit-specific and population-average results are to be expected:

- If all predictors are held constant at their means, and if their means are zero, the population-average intercept can be used to estimate the average probability of success across the entire population, that is

$$\hat{\phi}_{ij} = \frac{1}{1 + \exp(-\gamma_{00}^*)}$$

This will not be true of unit-specific intercepts unless the average probability of success is very close to .5.

- Coefficient estimates (other than the intercept) based on the population-average model will often tend to be similar to those based on the unit-specific model but will tend to be smaller in absolute value.

Users will need to take care in choosing unit-specific versus population-average results for their research. The choice will depend on the specific research questions that are of interest. In the previous example, if one were primarily interested in how a change in W can be expected to affect a particular individual school's mean, one would use the unit-specific model. If one were interested in how a change in W can be expected to affect the overall population mean, one would use the population-average model.

Final estimation of fixed effects: (Population-average model)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-1.748596	0.087745	-19.928	354	<0.001
MSESC, γ_{01}	-0.283515	0.184451	-1.537	354	0.125
For MALE slope, β_1					
INTRCPT2, γ_{10}	0.446558	0.067024	6.663	739	<0.001
For PPED slope, β_2					
INTRCPT2, γ_{20}	-0.536527	0.088442	-6.066	739	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-1.748596	0.174018	(0.146,0.207)
MSESC, γ_{01}	-0.283515	0.753132	(0.524,1.083)
For MALE slope, β_1			
INTRCPT2, γ_{10}	0.446558	1.562924	(1.370,1.783)
For PPED slope, β_2			
INTRCPT2, γ_{20}	-0.536527	0.584776	(0.492,0.696)

Final estimation of fixed effects (Population-average model with robust standard errors)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	-1.748596	0.082243	-21.261	354	<0.001
MSESC, γ_{01}	-0.283515	0.196226	-1.445	354	0.149
For MALE slope, β_1					
INTRCPT2, γ_{10}	0.446558	0.062859	7.104	739	<0.001
For PPED slope, β_2					
INTRCPT2, γ_{20}	-0.536527	0.082355	-6.515	739	<0.001

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-1.748596	0.174018	(0.148,0.205)
MSESC, γ_{01}	-0.283515	0.753132	(0.512,1.108)
For MALE slope, β_1			
INTRCPT2, γ_{10}	0.446558	1.562924	(1.381,1.768)
For PPED slope, β_2			
INTRCPT2, γ_{20}	-0.536527	0.584776	(0.497,0.687)

Substituting the unit-specific estimates obtained from the output file we can express the expected η as

$$\hat{\eta}_{ij} = \left[-1.748596 + 0.5446558MALE_{ij} - 0.536527PPED_{ij} \right]$$

We know that there are four groups of respondents: females with pre-primary education, females without pre-primary education, and males with or without pre-primary education.

The possible estimated values of $\hat{\eta}$ are then

$$\hat{\eta}_{ij}(\text{males}, \text{pped} = 1) = [-1.748596 + 0.5446558 - 0.536527]$$

$$\hat{\eta}_{ij}(\text{males}, \text{pped} = 0) = [-1.748596 + 0.5446558]$$

$$\hat{\eta}_{ij}(\text{females}, \text{pped} = 1) = [-1.748596 - 0.536527]$$

$$\hat{\eta}_{ij}(\text{females}, \text{pped} = 0) = [-1.748596]$$