

October 11, 2007
SC705: Advanced Statistics
Instructor: Natasha Sarkisian
Class notes: Non-continuous Dependent Variables

So far we've only dealt with continuous dependent variables, but HLM allows us to estimate models when the dependent variables are categorical. Your dependent variable can be dichotomous (0/1), categorical with multiple unordered categories, ordinal, or count variable. In such cases, linear models are inappropriate as there are no restrictions on the predicted values of level-1 outcome, the level-1 random effect (i.e. level 1 residual) cannot be normally distributed, and cannot have homogenous variance (the variance depends on the predicted value). Therefore, we need to use HGLM models for such variables. Like in non-hierarchical analysis, this is accomplished by specifying a link function that transforms the dependent variable so that the level-1 predicted values are constrained to be within a specific interval. Specifically, we use logit models for dichotomous variables, multinomial logit for categorical with unordered categories, ordered logit for ordinal variables, and Poisson models for count variables.

The following is an example of analysis with a dichotomous dependent variable. We'll use THAIUGRP.MDM in Examples/Chapter 6. These are data on 7,516 sixth graders nested within 356 primary schools from a national survey of primary education in Thailand, conducted in 1988. The dependent variable of interest is the probability that a child will repeat a grade during the primary years (REP1). The level-1 independent variables are whether a child attended pre-primary education (PPED) and child's gender (MALE). The level-2 variable is mean SES of school (MSESC). To specify that the dependent variable is binary, we go to the Basic Settings menu and select Bernoulli option.

LEVEL 1 MODEL

$$\text{Prob}(\text{REP1}_{ij}=1|\beta_j) = \varphi_{ij}$$

$$\text{Log}[\varphi_{ij}/(1 - \varphi_{ij})] = \eta_{ij}$$

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

LEVEL 2 MODEL

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{MSESC}_j - \overline{\text{MSESC}}) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{MSESC}_j - \overline{\text{MSESC}}) + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(\text{MSESC}_j - \overline{\text{MSESC}}) + u_{2j}$$

The model specified for the fixed effects was:

```
-----
Level-1                                Level-2
Coefficients                            Predictors
-----
      INTRCPT1, B0                      INTRCPT2, G00
$      MASESC, G01
      MALE slope, B1                    INTRCPT2, G10
$      MASESC, G11
      PPED slope, B2                    INTRCPT2, G20
$      MASESC, G21
'$' - This level-2 predictor has been centered around its grand mean.
```

Summary of the model specified (in equation format)

Level-1 Model

$$\text{Prob}(Y=1|B) = P$$

$$\log[P/(1-P)] = B_0 + B_1*(MALE) + B_2*(PPED)$$

Level-2 Model

$$B_0 = G_{00} + G_{01}*(MSESC) + U_0$$

$$B_1 = G_{10} + G_{11}*(MSESC) + U_1$$

$$B_2 = G_{20} + G_{21}*(MSESC) + U_2$$

$$\text{Level-1 variance} = 1/[P(1-P)]$$

RESULTS FOR NON-LINEAR MODEL WITH THE LOGIT LINK FUNCTION: Unit-Specific Model
(macro iteration 755)

Tau

INTRCPT1,B0	1.31543	0.06393	-0.21242
MALE,B1	0.06393	0.10023	0.04712
PPED,B2	-0.21242	0.04712	0.09011

Tau (as correlations)

INTRCPT1,B0	1.000	0.176	-0.617
MALE,B1	0.176	1.000	0.496
PPED,B2	-0.617	0.496	1.000

Random level-1 coefficient	Reliability estimate
INTRCPT1, B0	0.378
MALE, B1	0.047
PPED, B2	0.028

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

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	-2.043050	0.095356	-21.425	354	0.000
MSESC, G01	-0.410774	0.249833	-1.644	354	0.101
For MALE slope, B1					
INTRCPT2, G10	0.465559	0.076924	6.052	354	0.000
MSESC, G11	0.270760	0.199846	1.355	354	0.176
For PPED slope, B2					
INTRCPT2, G20	-0.532227	0.097716	-5.447	354	0.000
MSESC, G21	-0.044859	0.253619	-0.177	354	0.860

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, B0			
INTRCPT2, G00	-2.043050	0.129633	(0.107,0.156)
MSESC, G01	-0.410774	0.663137	(0.406,1.083)
For MALE slope, B1			
INTRCPT2, G10	0.465559	1.592905	(1.370,1.853)
MSESC, G11	0.270760	1.310960	(0.885,1.941)
For PPED slope, B2			
INTRCPT2, G20	-0.532227	0.587296	(0.485,0.712)
MSESC, G21	-0.044859	0.956132	(0.581,1.573)

The outcome variable is REP1
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, B0					
INTRCPT2, G00	-2.043050	0.094461	-21.628	354	0.000
MSESC, G01	-0.410774	0.255098	-1.610	354	0.108
For MALE slope, B1					
INTRCPT2, G10	0.465559	0.075695	6.150	354	0.000
MSESC, G11	0.270760	0.204300	1.325	354	0.186
For PPED slope, B2					
INTRCPT2, G20	-0.532227	0.095718	-5.560	354	0.000
MSESC, G21	-0.044859	0.253801	-0.177	354	0.860

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, B0			
INTRCPT2, G00	-2.043050	0.129633	(0.108,0.156)
MSESC, G01	-0.410774	0.663137	(0.402,1.094)
For MALE slope, B1			
INTRCPT2, G10	0.465559	1.592905	(1.373,1.848)
MSESC, G11	0.270760	1.310960	(0.878,1.958)
For PPED slope, B2			
INTRCPT2, G20	-0.532227	0.587296	(0.487,0.709)
MSESC, G21	-0.044859	0.956132	(0.581,1.574)

Final estimation of variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, U0	1.14692	1.31543	237	423.83711	0.000
MALE slope, U1	0.31659	0.10023	237	213.85874	>.500
PPED slope, U2	0.30018	0.09011	237	166.92266	>.500

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

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	-1.671706	0.082783	-20.194	354	0.000
MSESC, G01	-0.398163	0.221047	-1.801	354	0.072
For MALE slope, B1					
INTRCPT2, G10	0.426803	0.060136	7.097	354	0.000
MSESC, G11	0.248359	0.162132	1.532	354	0.126
For PPED slope, B2					
INTRCPT2, G20	-0.481299	0.079160	-6.080	354	0.000
MSESC, G21	-0.000757	0.215165	-0.004	354	0.997

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, B0			
INTRCPT2, G00	-1.671706	0.187926	(0.160,0.221)
MSESC, G01	-0.398163	0.671553	(0.435,1.037)
For MALE slope, B1			
INTRCPT2, G10	0.426803	1.532350	(1.362,1.724)
MSESC, G11	0.248359	1.281920	(0.932,1.763)
For PPED slope, B2			

INTRCPT2, G20	-0.481299	0.617980	(0.529,0.722)
MSESC, G21	-0.000757	0.999243	(0.655,1.525)

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, B0					
INTRCPT2, G00	-1.671706	0.059774	-27.967	354	0.000
MSESC, G01	-0.398163	0.170459	-2.336	354	0.020
For MALE slope, B1					
INTRCPT2, G10	0.426803	0.045573	9.365	354	0.000
MSESC, G11	0.248359	0.133309	1.863	354	0.063
For PPED slope, B2					
INTRCPT2, G20	-0.481299	0.059564	-8.080	354	0.000
MSESC, G21	-0.000757	0.173800	-0.004	354	0.997

Fixed Effect	Coefficient	Odds Ratio	Confidence Interval
For INTRCPT1, B0			
INTRCPT2, G00	-1.671706	0.187926	(0.167,0.211)
MSESC, G01	-0.398163	0.671553	(0.480,0.939)
For MALE slope, B1			
INTRCPT2, G10	0.426803	1.532350	(1.401,1.676)
MSESC, G11	0.248359	1.281920	(0.987,1.666)
For PPED slope, B2			
INTRCPT2, G20	-0.481299	0.617980	(0.550,0.695)
MSESC, G21	-0.000757	0.999243	(0.710,1.406)

Variance components

Note that the variance component does not contain an estimate of level 1 variance. That is because in logistic regression models, it is not possible to estimate both the coefficients and the error variance; therefore, in all logistic regression models, the error variance is always fixed to the same number which is $\pi^2/3 = 3.29$. That rule also applies to multilevel models, but only to their level 1 residuals. Knowing this means that we can calculate the intraclass correlation coefficient or the proportion of variance explained. For both, we can follow the procedures described on pp.224-227 of the Snijders and Bosker chapter on dichotomous outcomes. For instance, the ICC would be calculated as

$$\rho_1 = \frac{\tau_0^2}{\tau_0^2 + \pi^2/3}$$

And the proportion of variance explained can be calculated as

$$R_{\text{dicho}}^2 = \frac{\sigma_F^2}{\sigma_F^2 + \tau_0^2 + \sigma_R^2}$$

Note that in addition to the level 2 intercept variance τ_0 and level 1 variance $\sigma_R^2 = 3.29$, we need to know the variance of fitted values σ_F^2 . That refers to the variance of linear predictions, which are the values that results if we multiply our coefficients by our variable values and add up these products. That is, we are talking about the predicted values of logits. To obtain the variance of fitted values, we can use level 1 residuals file and calculate the variance of the FITVAL variable containing the linear predictor values. Note that such R squared values are typically lower than values we are used to with OLS because σ_R^2 is a fixed number.

Unit-specific versus population-average models

The distinction between unit-specific and population-average models emerges when we use nonlinear link models (HGLM). The unit-specific model presents coefficients for a hypothetical unit (group) where random effect is zero. The population-average model presents coefficients averaged out for the whole sample.

Further, these two models make different assumptions about the underlying distribution of random effects and they are oriented towards different research aims. The unit-specific models are more appropriate for describing how the effects of level 1 and level 2 predictors vary across level 2 units. Population-average models, in contrast, give answers to population-average questions – that is, they are more appropriate for estimating predicted probabilities for the whole population. If we use a regression model to examine how preprimary education experience relates to the risk of class repetition in different schools, we are asking a unit-specific question. If we want to know how the risk of repetition differs between those who do and do not have preprimary experience nationwide, we need a population-average estimate.

Also note that population-average inferences are based on fewer assumptions than unit-specific inferences and are therefore more robust to erroneous assumptions about the random effects in the model. In a way, unit-specific models are richer but more sensitive to model assumptions.

In regular HLM, these are the same, but in HGLM they differ because nonlinear transformations, such as that from probability into log odds, mean that the distribution of predicted probabilities is not symmetric. The two estimates become rather similar when the fixed effect is close to 0 or the random component is close to 0.

Interpreting fixed effects

The interpretation of the fixed effects is very similar to the interpretation of the results of logistic regression—but be careful as we still have variables on multiple levels and can potentially have interactions across levels. Interpreting coefficients themselves allows us to discuss the direction and significance of effects, but not their size. To talk about the size, we use odds ratios. When interpreting odds ratios, do not forget that these are multiplicative coefficients, so if you want to interpret, for example, an interaction term, you would have to multiply rather than add the odds ratio numbers. Alternatively, you can add the numbers presented in the coefficient column and then exponentiate the result.

In addition, we can calculate predicted probabilities (P) by calculating predicted logits (L) and then recalculating them into probabilities

Since $L = \log(\text{odds}) = \log(P/(1-P))$, then

$$P = e^L / (1 + e^L)$$

As mentioned above, predicted logits L are available in the FITVAL variable, so you can easily generate predicted probabilities on the basis of that. It is more interesting for interpretation purposes, however, to calculate predicted probabilities for some hypothetical, strategically selected cases. For that, you have to calculate the logit of interest by hand by plugging values into the equation:

$$L = \gamma_{00} + \gamma_{01} * \text{MSESC} + \gamma_{10} * \text{MALE} + \gamma_{11} * \text{MALE} * \text{MSESC} + \gamma_{20} * \text{PPED} + \gamma_{21} * \text{PPED} * \text{MSESC}$$

For instance, if we want to calculate the predicted probabilities for males and females who did not attend pre-primary education and who are in a school with average SES, we calculate:

$$L \text{ for males} = \gamma_{00} + \gamma_{10} * \text{MALE} = -1.67 + 0.43 = -1.24$$

$$L \text{ for females} = \gamma_{00} = -1.67$$

Corresponding probabilities would be:

$$P \text{ for males} = \frac{\exp(-1.24)}{1 + \exp(-1.24)} = .22$$

$$P \text{ for females} = \frac{\exp(-1.67)}{1 + \exp(-1.67)} = .16$$

Such strategically calculated predicted probabilities are very useful for a more intuitive presentation of results. Note, however, that the differences between probabilities are not constant – e.g. if we looked at those who attended preprimary education (PPED=1), the difference between males and females would change.

We can also calculate predicted probabilities for specific schools – we calculate linear predictors and add the corresponding level 2 residual. For example, we take female students with no preprimary education in schools 10103 and 10104. We would need to know MSESC and EBINTRCPT1 for their schools. Since we used MSESC as a mean-centered variable, we need to get its values based on that; for that, we generate a mean-centered variable in our level 2 residuals file:

```
. sum MSESC
```

Variable	Obs	Mean	Std. Dev.	Min	Max
MSESC	356	.0078371	.3806623	-.77	1.49

```
. gen MSESCm = MSESC-.0078371
```

School	N	MSESCm	L for MALE=0 PPED=0 ($L = \gamma_{00} + \gamma_{01} * \text{MSESC}$)	Predicted probability based on fixed effects only	Random effect EBINTRCPT1	Predicted probability based on fixed and random effects
10103	17	.87	$-1.67 - 0.4 * .87 = -2.018$.117	-.216	.097
10104	29	.19	$-1.67 - 0.4 * .19 = -1.746$.149	-1.228	.049

Thus we see that school 10104 has advantages because of its mean SES, but its unique component suggests that it is much less efficient than school 10103 – in fact the unique component of school 10104 more than compensates for its economic disadvantage.

Note that if we focused on male students, we'd have to take into account both the fixed effect (coefficient) for MALE and the value of random effect for the MALE slope which is stored in EBMALE.

Example: Litwin, Kenneth J. 2004. "A Multilevel Multivariate Analysis of Factors Affecting Homicide Clearances." *Journal of Research in Crime and Delinquency*, 41, 327-351.

1. What are the research questions in this study?

2. How do the authors justify the use of HGLM?

3. What are the two levels?

4. What variables are used on level 1?

5. What variables are used on level 2?

6. Interpret the results presented in Table 1. What are the main findings?

7. In addition to what the author presents, how else could he present his results?

8. What information about the estimated model is missing from this article?