Generalized Multilevel Linear Models

Introduction to Multilevel
Models Workshop

University of Georgia:

Institute for Interdisciplinary Research in
Education and Human Development

Hierarchical Generalized Linear Models

- Introduction to generalized models
  - Models for binary outcomes
  - Interpreting parameter estimates
- Fun with estimation
- Wrapping up…

A Taxonomy of Linear Models

- Linear statistical models can be organized by:
  - Type of outcome:
    - Normal (general) versus Non-Normal (general*ized*)
  - Type of sampling:
    - Simple/Independent (one error term)
    - Complex/Dependent (multiple error terms)

- The term "linear" loosely describes the relationship between the predictors and the response variable
  - Non-normal response variables are *link-transformed*
  - Link-transformed: modeled using a function of the variable

Naming Conventions of Models

- The names of models:
  - General Linear Models:
    - fixed effects/ NO random effects / NO link function
  - General Linear Mixed Models:
    - fixed AND random effects / NO link function
  - General*ized* Linear Models:
    - fixed effects / NO random effects / link function
  - General*ized* Linear Mixed Models:
    - fixed AND random effects with link function

Binary versus Continuous Outcome Variables

- Variable types:
  - Continuous: ranges from negative infinity to infinity
  - Binary: 0/1
- Means:
  - Continuous outcome mean: $\bar{y}$
  - Binary outcome mean: proportion of 1's = $\pi_y \rightarrow p_y$
- Variances:
  - Continuous: $\text{Var}(Y) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$
  - Binary: $\text{Var}(Y) = p_y(1-p_y) = p_y q_y = \sigma_y^2 \rightarrow s_y^2$
    - **The variance IS determined by the mean!**

TABLE 3.2
Binary Item Variance and Difficulty

p	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
variance	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

A General Linear Model for Binary Outcomes

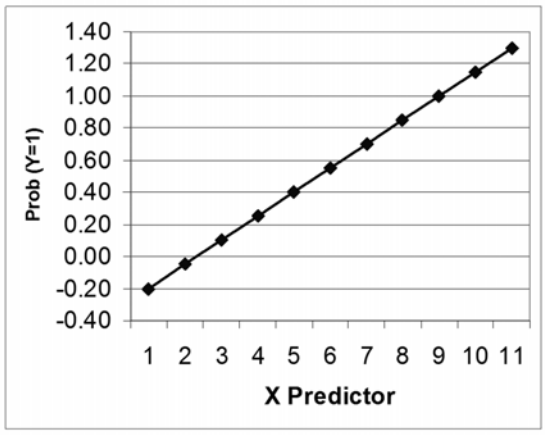
- If your outcome variable is binary (0 or 1):
  - Expected mean is proportion of people who have a 1 (or "p", the probability of Y=1)
    - The probability of having a 1 is what we're trying to predict for each person, given the values on the predictors
- Under the general linear model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$
  - β_0 = expected probability when all predictors are 0
  - β's = expected change in probability for a one-unit change in the predictor
  - e_i = difference between observed and predicted values
- Model becomes Y_i = (predicted probability of 1) + e_i

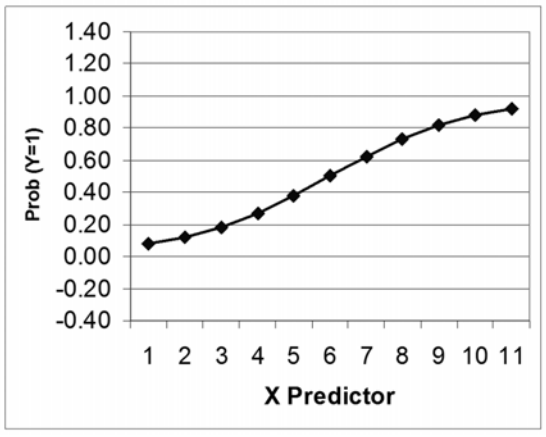
A General Linear Model for Binary Outcomes

- But if Y_i is binary, then e_i can only be two things:
  - e_i = Observed Y_s minus Predicted Y_s
    - If $Y_i = 0$ then e = (0 − predicted probability)
    - If $Y_i = 1$ then e = (1 − predicted probability)
- Mean of errors would still be 0…
- Variance of errors cannot be constant over levels of X like we assume in general linear models
  - The mean and variance of a binary outcome are dependent
  - This means that because the conditional mean of Y (p, the predicted probability Y= 1) is dependent on X, *then so is the error variance*

A General Linear Model for Binary Outcomes

- Needed: a method to translate probabilities bounded by zero and one to the entire number line
- Options:
  - Ignore bounding and use traditional general linear model
  - Transform probability to something continuous





3 Problems with General Linear Models for Binary Outcomes

1. Restricted range (e.g., 0 to 1 for binary item)
   - Predicted values can each only be off in two ways
     → So residuals can't be normally distributed

2. Variance is dependent on the mean, and not estimated
   - Fixed and random parts are related
     → So residuals can't have constant variance

3. Residuals have a limited number of possible values
   - Predicted values can each only be off in two ways
     → So residuals can't be normally distributed

Differing Types of Outcomes

- **General*ized* Linear Models** are General Linear Models
  - with differently distributed error terms
  - with transformed outcome variables

- Many kinds of non-normally distributed outcomes have some kind of generalized linear model to go with them:
  - **Binary (dichotomous)**
  - Unordered categorical (nominal)
  - Ordered categorical (ordinal)
  - Counts (discrete, positive values)

  } These two are often called "multinomial" inconsistently

  - Censored (piled up and cut off at one end – left or right)
  - Zero-inflated (pile of 0's, then some distribution after)

Parts of a General*ized* Linear Model

- Link Function (main difference from GLM):
  - How a non-normal outcome gets transformed into something that is continuous (unbounded)
  - For outcomes that are already normal, general linear models are just a special case with an "identity" link function (Y * 1)

- Fixed Effects:
  - How predictors linearly relate to the transformed outcome
  - New transformed $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$

- Random Effects (or Error Variances):
  - Level-1 errors aren't normal and homoscedastic
  - Family of alternative distributions at our disposal that map onto what the distribution of errors could possibly look like

General*ized* Models for Binary Outcomes

- Rather than modeling the probability of a 1 directly, we need to transform it into a more continuous variable with a link function, for example:

  - Transform probability into an odds ratio:
    - Odds ratio: (p / 1-p) = prob(1) / prob(0)
    - If p = .7, then Odds(1) = 2.33; Odds(0) = .429
    - Odds scale is way skewed, asymmetric, and ranges from 0 to infinity
  - Take *natural log of odds ratio* : called "logit" link
    - LN (p / 1-p): Natural log of (prob(1) / prob(0))
    - If p = .7, then LN(Odds(1)) = .846; LN(Odds(0)) = -.846
    - Logit scale is now symmetric about 0

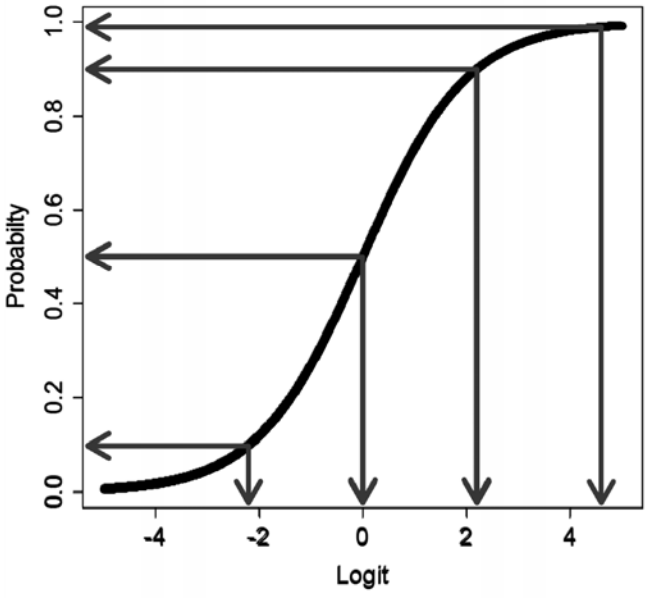
Model Background

- The log-odds is called a ***logit***

$$\text{Logit}\left(P\left(Y=1\right)\right)=\ln\left(\frac{P\left(Y=1\right)}{1-P\left(Y=1\right)}\right)$$

- The logit is used because the responses are binary
  - Responses are either (1) or (0)

More on Logits



Probability	Logit
0.5	0.0
0.9	2.2
0.1	-2.2
0.99	4.6

From Logits to Probabilities

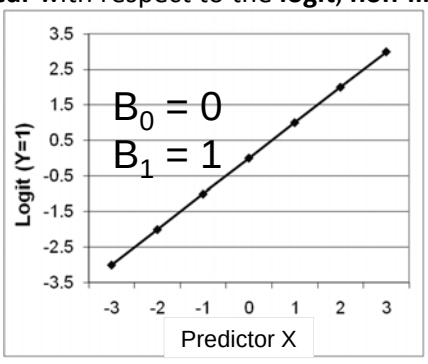
- Whereas logits are useful as the are unbounded continuous variables, categorical data analyses rely on estimated probabilities

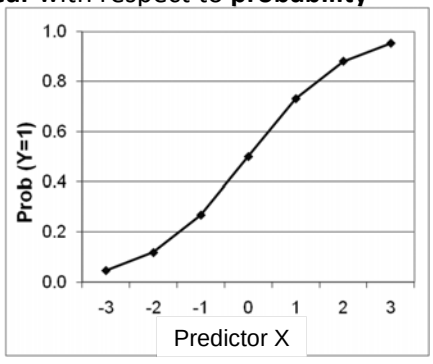
- The inverse logit function coverts the unbounded logit to a probability
  - This is also the form of an IRT model (and logistic regression)

$$P(Y=1)=\frac{\exp(\text{Logit}(P(Y=1)))}{1+\exp(\text{Logit}(P(Y=1)))}$$

Non-Linearity in Prediction

- The relationship between X and the probability of response=1 is "non-linear" → an s-shaped logistic curve whose shape and location are dictated by the estimated fixed effects
  - **Linear** with respect to the **logit**, **non-linear** with respect to **probability**





- The logit version of the model will be easier to explain; the probability version of the prediction will be easier to show.

The Logistic Model

- Outcome is log odds (logit) of probability instead of probability
  - Symmetric, unbounded outcome
  - Assume linear relationship between predictors and log odds (logit)
  - This allows an overall non-linear (S-shaped) relationship between X's and probability of Y=1

- Errors are ***not*** assumed to be normal with constant variance
  - 'e_i' will be missing – residual variance is NOT estimated
  - Errors are assumed to follow a logistic distribution with a known residual variance of $\pi^2/3$ (3.29)
  - Still assume errors are independent
    - Clustered data would need a generalized *mixed* model that would include random effects that account for any dependency

Model Estimation and Comparison

- Not in least squares anymore…done with ML
  - Tries to maximize likelihood of observed frequency of discrete responses for Y given model parameters
  - No REML in generalized models
- Model comparisons follow the rules you already know:
  - Nested models can be compared with deviance difference tests
    - Difference compared to χ^2 with df = difference in # parameters
  - Non-nested models can be compared with AIC and BIC
    - Smaller is better, no critical values, so no significance tests
- No single version of R^2 for generalized models
  - Because variance in outcome depends on proportion of 1's, more lop-sided outcomes will have less variance to account for
  - Several approximations are available

Our New Model using LN(Odds):

- Log Odds model: $LN(p/1\text{-}p)_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$
- Because the model for the means is the same as in general regression, any kind of predictor can be included
  - Dummy codes, continuous variables, and their interactions
- Regardless of which form (probability, odds, logit), effects of predictors will be monotonic
- Predictor effects are linear and additive like in regression, but what does a 'change in the logit' mean anyway?

Example: Categorical Predictor

DV = "Admission" (0=no, 1=yes)
IV = Sex (0=F, 1=M)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	Gender	1.695	.976	3.015	1	.082	5.444
	Constant	-.847	.690	1.508	1	.220	.429

1. Variable(s) entered on step 1: Gender.

- Logit: $LN(p/1\text{-}p)_i = \beta_0 + \beta_1 x_i$
  - Logit Y_s = -.847 + 1.695(M_i): note is additive
  - Log odds for women = -.847, for men = .847
- Odds: $(p/1\text{-}p)_i = \exp(\beta_0) * \exp(\beta_1 x_i)$
  - Odds Y_i = exp(-.847) * exp(1.695(M_i=1)): note is multiplicative
  - Multiply, then exp: Odds Y_i= .429 * 5.444 = .429 for W, 2.34 for M
- Prob: $P(1)_i = \exp(\beta_0) * \exp(\beta_1 x_i) / 1 + (\exp(\beta_0) * \exp(\beta_1 x_i))$
- $Prob(Y=1)_i$ = (.429 * 5.444) / 1 + (.429 * 5.444) = .70
- So, for men, probability(admit) = .70, odds = 2.34, log odds = .847
  for women, probability(admit) = .30, odds = .429, log odds = -.847

Example: Continuous Predictor

DV = "Senility"
(0=no, 1=yes)
IV = WAIS (0=10)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[1]	WAIS	-.275	.103	7.120	1	.008	.760
	Constant	1.776	1.067	2.771	1	.096	5.907

1. Variable(s) entered on step 1: WAIS.

- Logit: $LN(p/1-p)_i = \beta_0 + \beta_1 x_i$
  - Logit Y_i = 1.776 - .275($WAIS_i$=1) = 1.50: note is additive
  - For every one-unit change in WAIS, log odds go down by .275
- Odds: $(p/1-p)_i = \exp(\beta_0 + \beta_1 x_i)$ or $\exp(\beta_0) * \exp(\beta_1 x_i)$
  - Odds Y_i = exp(1.776) * exp(- .275($WAIS_i$=1)): note is multiplicative
  - Multiply, then exp: Odds Y_i = 5.907 * .760 = 4.49
- Prob: $P(1)_i = \exp(\beta_0 + \beta_1 x_i) / 1 + \exp(\beta_0 + \beta_1 x_i)$
  - Prob(Y=1)$_i$ = 5.907 * .760($WAIS_i$=1) / 1 + 5.907 * .760($WAIS_i$=1) = .82
- So, if WAIS=10, prob(senility) = .86, odds = 5.91, log odds = 1.78
- So, if WAIS=11, prob(senility) = .82, odds = 4.49, log odds = 1.50
- So, if WAIS=16, prob(senility) = .53, odds = 1.13, log odds = 0.13
- So, if WAIS=17, prob(senility) = .46, odds = 0.86, log odds = -0.15

Generalized (Linear) Mixed Models for Clustered Discrete Outcomes

- Same components as generalized models:
  - Link function to transform outcome variable into some continuous
  - Linear predictive model (linear for link-transformed outcome variable)
  - Alternative distribution of errors assumed
- The difference is that we will add random effects to address dependency in clustered data
  - The big difference is that variances are ADDED TO, not EXTRACTED FROM, the original residual variance
  - Thus, some concepts translate exactly from general linear mixed models, but some don't

Empty Logistic Mixed Model

- Level 1: $\text{Logit}(Y_{is}) = B_{0s}$
- Level 2: $B_{0s} = \gamma_{00} + U_{0s}$

Note what's NOT in level 1…

- Combined: $\text{Logit}(Y_{is}) = \gamma_{00} + U_{0s}$
- Residual variance is not estimated : $\pi^2/3$, or 3.29
  - (Known) residual is in model for actual Y, not prob(Y) or logit(Y)
- Logistic ICC = $\dfrac{\text{Var}(U_{0s})}{\text{Var}(U_{0s}) + \text{Var}(e_{is})} \quad \dfrac{\text{Var}(U_{0s})}{\text{Var}(U_{0s}) + 3.29}$
- Can do ML deviance difference test to see if $\text{Var}(U_{0s}) > 0$

Logistic Mixed Model: Random Intercepts and Slopes

- Level 1: $\text{Logit}(Y_{is}) = B_{0s} + B_{1s}x_{is}$
- Level 2: $B_{0s} = \gamma_{00} + U_{0s}$
  $B_{1s} = \gamma_{10} + U_{1s}$
- Combined: $\text{Logit}(Y_{is}) = (\gamma_{00} + U_{0s}) + (\gamma_{10} + U_{1s})(x_{is})$
- Residual variance is still not estimated: $\pi^2/3$, or 3.29
- Can test new fixed or random effects with ML deviance differences tests (or Wald test p-values for fixed effects)

New Interpretation of Fixed Effects

- In general linear mixed models, the fixed effects are interpreted as the 'average' effect for the sample
  - γ_{00} is 'sample average' intercept
  - U_{0i} is 'individual deviation from sample average'
- What 'average' means in general*ized* linear mixed models is different, because the natural log is a nonlinear function:
  - So the mean of the logs ≠ log of the means
  - Therefore, the fixed effects are not the 'sample average' effect, they are the effect for someone *specifically with a* $U_i = 0$
    - Fixed effects are *conditional* on the random effects
    - This gets called a "unit-specific" or "subject-specific" model
    - This distinction does not exist for normally distributed outcomes

New Comparisons across Models

- NEW RULE: Coefficients cannot be compared across models, because they are not on the same scale!
- Two reasons for this, both due to residual variance = 3.29:
  - When adding a random intercept to an empty model, the total variation in the outcome has increased
  - the fixed effects will increase in size because they are *unstandardized* slopes
  - Residual variance can't decrease due to effects of level-1 predictors, so all other estimates have to go up to compensate

$$\gamma_{mixed} \approx \sqrt{\frac{Var(U_{0i})+3.29}{3.29}}(B_{fixed})$$

Pseudo-R^2: Even more 'Psuedo' than before…

- The "total variance of Y" is a fluid concept depending on how many variance components and fixed effects are in the model

- Psuedo-R^2 is calculated in two steps:
  - Calculate all predicted y's, get their variance
  - R^2 = "explained" variance / "total" variance
  - $R^2 = Var(y_{pred}) / Var(y_{pred}) + Var(U_{0i}) + 3.29$
    - Nothing in texts about what to do with random slope variance

A Little Bit about Estimation

- Goal: End up with maximum likelihood estimates for all model parameters (because they are consistent, efficient)
  - When model variances are normal at level 1 and normal at level 2, this is relatively easy to do
  - When model variances are non-normal at level 1 but normal at level 2, this is much harder to do
- Bottom line: Estimating models on discrete outcomes is much harder than for normally distributed outcomes, so some kind of approximation is usually necessary
- 2 main families of approaches:
  - Quasi-Likelihood methods ("marginal/penalized quasi ML": SAS proc glimmix)
  - Numerical Integration ("adaptive Gaussian quadrature": SAS proc nlmixed)
  - Also Bayesian methods (available in SAS proc mcmc)

2 Families of Estimation

- Quasi-Likelihood methods → in PROC GLIMMIX (for 3+ levels)
  - "Marginal QL" → approximation around fixed part of model
  - "Penalized QL" → approximation around fixed + random parts
  - These both underestimate variances (MQL more so than PQL)
  - 2nd-order QL is supposed to be better than 1st-order QL
  - QL methods DO NOT PERMIT MODEL DEVIANCE COMPARISONS
  - HLM program adds Laplace approximation to QL, which then does permit deviance comparisons

- Numerical Integration → in Mplus and PROC NLMIXED (2 levels)
  - Better estimates, but can take for-freaking-ever
  - DOES permit regular model deviance comparisons
  - Will blow up with many random effects (which make the model exponentially more complex)
  - Good idea to use QL to get start values first to then use in integration

Summary: Differences in General*ized* Mixed Models

- Analyze link-transformed DV
  - Linear relationship between X's and transformed continuous Y
  - Non-linear relationship between X's and original discrete Y
- Residual variance is not estimated
  - So it can't go down after adding level-1 predictors
  - So the scale of everything else has to go up to compensate
  - Scale will also be different after adding random effects for the same reason – the total variation in the model is now bigger
  - Fixed effects may not be comparable across models as a result
- Estimation is trickier and takes longer
  - Numerical integration is best but may blow up in complex models
  - Start values are often essential (can get those with QL estimator)

GENERALIZED LINEAR MIXED MODELS EXAMPLE #1

Practicing What We've Learned

- In order to practice what we have learned, let's run an actual example and see if we can predict a student's SES
  - In our data, this is free/reduced price lunch
- Data come from the end-of-grade mathematics exam from a "rectangular Midwestern state"
  - 94 schools
  - 13,804 students (schools had between 31 and 515 students)
- Variables of interest:
  - frlunch = free/reduced price lunch code (0=no; 1=F/R lunch)
  - math = score on EOG math exam (ranging from 0 to 83)
  - boyvsgirl = gender code (boy=0; girl=1)
  - nonwhite = ethnicity code (white = 0; non-white = 1)
  - schoolID = school ID number

The Data…A Summary

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
math	math: Math Test Score Outcome	13082	48.1185599	17.2590473	0	83.0000000
SMCmath	SMCmath: School Mean Centered Math)	13082	-3.91065E-17	15.8552141	-44.6506986	44.8818898
SMmath	SMmath: School Mean Math	13082	48.1185599	6.8181301	29.4509804	61.6136364
boyvsgirl	boyvsgirl: Boy=0, Girl=1	13082	0.4981654	0.5000157	0	1.0000000
frlunch	frlunch: 0=No, 1=Free/Reduced Lunch	13082	0.3075218	0.4614850	0	1.0000000
nonwhite	nonwhite: White=0, NonWhite=1	13082	0.3049992	0.4604247	0	1.0000000
SchoolN	SchoolN: # Students Contributing Data	13082	274.9501605	155.3319041	31.0000000	515.0000000

Here: Y = frlunch

Model 1A: The Empty Model

- Model:

$$Logit\big(P(Y_{is} = 1)\big) = \beta_0 + e_{is}$$

- Where $e_{is} \sim Logistic(0, \frac{\pi^2}{3})$

Fit Statistics

-2 Log Likelihood	16145.89
AIC (smaller is better)	16147.89
AICC (smaller is better)	16147.89
BIC (smaller is better)	16155.37
CAIC (smaller is better)	16156.37
HQIC (smaller is better)	16150.39
Pearson Chi-Square	13082.00
Pearson Chi-Square / DF	1.00

Parameter Estimates

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-0.8117	0.01895	13081	-42.84	<.0001

Interpreting Parameters

- The intercept parameter is the baseline logit
  - In a general linear model, this would be the grand mean

- $\hat{\beta}_0 = -0.812\ (0.019)$

- We can convert this back to a probability:

$$\frac{\exp(-.812)}{1+\exp(-.812)} = .307$$

- This is the overall proportion of students on free/reduced price lunch
  - The grand mean…

Model 1B: Adding a Random Intercept

- Level 1:

$$Logit\big(P(Y_{is}=1)\big) = \beta_{0s} + e_{is}$$

- Level 2:

$$\beta_{0s} = \gamma_{00} + U_{0s}$$

- Combined Model:

$$Logit\big(P(Y_{is}=1)\big) = \gamma_{00} + U_{0s} + e_{is}$$

- Where $e_{is} \sim Logistic(0, \frac{\pi^2}{3})$; $U_{0s} \sim N(0, \tau_0^2)$

Model 1B: Results

- Model Fit:

Fit Statistics

-2 Log Likelihood	13173.52
AIC (smaller is better)	13177.52
AICC (smaller is better)	13177.52
BIC (smaller is better)	13182.61
CAIC (smaller is better)	13184.61
HQIC (smaller is better)	13179.58

- Estimates:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	schoolID	1.9434	0.3289

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-1.1717	0.1490	93	-7.86	<.0001

- Notice how the intercept now changes
  - This didn't happen in the general linear mixed model

Which Model is Preferred?

- Luckily, SAS proc glimmix now provides an accurate test of the hypothesis that the random intercept variance is zero using a mixture chi-square

Tests of Covariance Parameters
Based on the Likelihood

Label	DF	-2 Log Like	ChiSq	Pr > ChiSq	Note
H0: Random Intercept Variance=0	1	16146	2972.37	<.0001	MI

MI: P-value based on a mixture of chi-squares.

- The p-value is small – we reject the null hypothesis
  - We need the random intercept

Model Summary

- Up next, we should describe how much dependency is present in our data
  - Harder to do in categorical data
    - No sufficient summary statistic exists

- We can form our estimated ICC (using 3.29 – the level one error variance):

$$\rho = \frac{1.9434}{1.9434 + 3.29} = 0.371$$

Model 2: Adding Continuous Predictors

- Now that we know what we have to work with, we can start adding predictors
  - We will start with our math score variable
  - Note: this is not the assumed causal order (being bad at math does not cause a student to need free or reduced lunch)

- Math, as a continuous variable, should be cluster mean centered so as to disentangle the effects at the varying levels
  - We will add both level 1 (cluster mean centered) and level 2 (school mean) to the analysis simultaneously

Model 2A: Adding Math

- Level 1:

$$Logit\big(P(Y_{is}=1)\big) = \beta_{0s} + \beta_{1s}(X_{is} - \bar{X}_s) + e_{is}$$

- Level 2:

$$\beta_{0s} = \gamma_{00} + \gamma_{01}\big(\bar{X}_s - \bar{X}_g\big) + U_{0s}$$
$$\beta_{1s} = \gamma_{10}$$

- Combined Model:

$$Logit\big(P(Y_{is}=1)\big) = \gamma_{00} + \gamma_{10}(X_{is} - \bar{X}_s) + \gamma_{01}\big(\bar{X}_s - \bar{X}_g\big) + U_{0s} + e_{is}$$

- Where $e_{is} \sim Logistic(0, \frac{\pi^2}{3})$; $U_{0s} \sim N(0, \tau_0^2)$

Model 2A: Results

- Model Fit:

-2 Log Likelihood	12391.37
AIC (smaller is better)	12399.37
AICC (smaller is better)	12399.38
BIC (smaller is better)	12409.55
CAIC (smaller is better)	12413.55
HQIC (smaller is better)	12403.48

- Estimates:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	schoolID	0.8361	0.1563

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-1.2740	0.1021	93	-12.48	<.0001
SMCmath	-0.03720	0.001450	12986	-25.66	<.0001
SMmath48	-0.1517	0.01463	12986	-10.37	<.0001

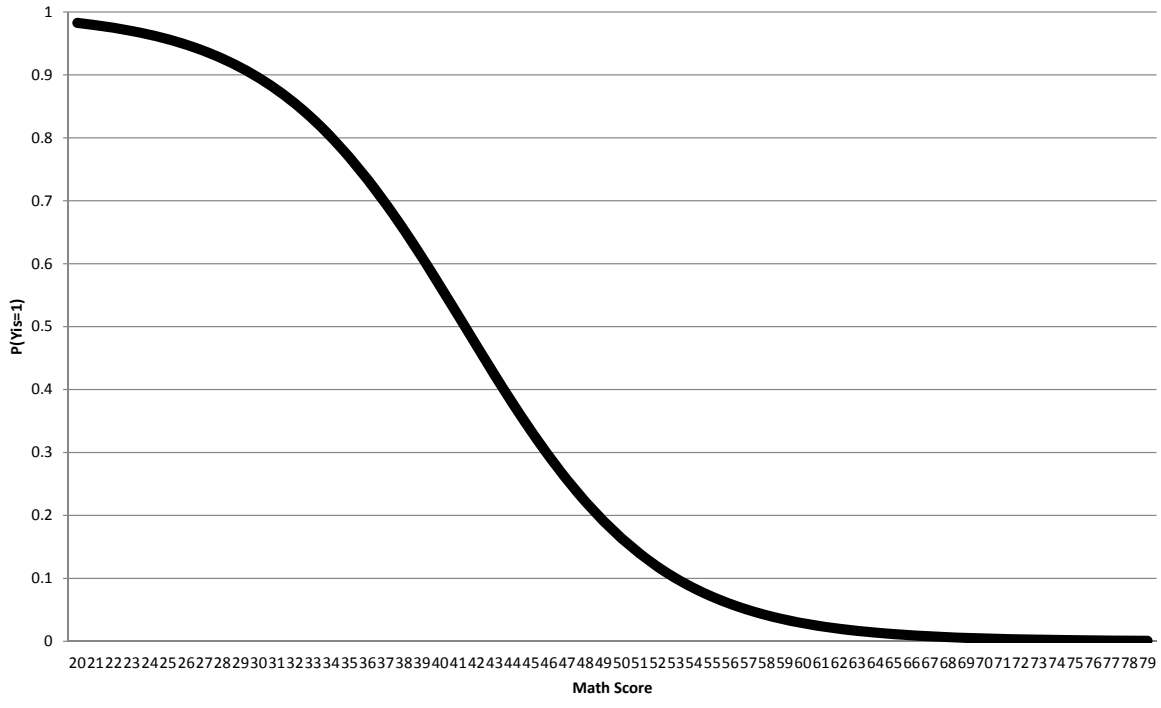
Which Model is Preferred?

- Because we are testing fixed effects, we must form a deviance test by hand
  - Model 1B -2LL: 13,173.52
  - Model 2A -2LL: 12,391.37

- Deviance test: 13,173.52 – 12,391.37 = 782.16
- df = 2
- P-value < 0.0001

- The p-value is small – we reject the null hypothesis
  - Model 2A is preferred…

Plot of Prediction of Free/Reduced Lunch



Model Summary

- Up next, we should describe how much dependency is present in our data

- We can form our estimated ICC (using 3.29 – the level one error variance):

$$\rho = \frac{.8361}{.8361 + 3.29} = 0.203$$

- We can also calculate our Pseudo-R^2:
  - 1.9434-.8361/1.9434 = 0.570

Model 2B: Adding a Random Slope

- Level 1:

$$Logit\big(P(Y_{is} = 1)\big) = \beta_{0s} + \beta_{1s}(X_{is} - \bar{X}_s) + e_{is}$$

- Level 2:

$$\beta_{0s} = \gamma_{00} + \gamma_{01}\big(\bar{X}_s - \bar{X}_g\big) + U_{0s}$$
$$\beta_{1s} = \gamma_{10} + U_{1s}$$

- Combined Model:

$$Logit\big(P(Y_{is} = 1)\big) = \gamma_{00} + \gamma_{10}(X_{is} - \bar{X}_s) + \gamma_{01}\big(\bar{X}_s - \bar{X}_g\big) + U_{0s} + U_{1s}(X_{is} - \bar{X}_s) + e_{is}$$

- Where $e_{is} \sim Logistic(0, \frac{\pi^2}{3})$; $[U_{0s,}U_{1s}] \sim N_2(\mathbf{0}, \boldsymbol{\tau})$

Model 2B: Results

- Model Fit:

Fit Statistics

-2 Log Likelihood	12353.04
AIC (smaller is better)	12365.04
AICC (smaller is better)	12365.04
BIC (smaller is better)	12380.30
CAIC (smaller is better)	12386.30
HQIC (smaller is better)	12371.20

- Estimates:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	schoolID	0.8058	0.1526
UN(2,1)	schoolID	-0.00348	0.002876
UN(2,2)	schoolID	0.000158	0.000055

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-1.2724	0.1007	93	-12.64	<.0001
SMCmath	-0.03436	0.002417	93	-14.21	<.0001
SMmath48	-0.1561	0.01473	1893	-10.60	<.0001

Model 2B: Testing for Random Slope

- Glimmix will do this for us – and we find that we need the random slope

Tests of Covariance Parameters Based on the Likelihood

Label	DF	-2 Log Like	ChiSq	Pr > ChiSq	Note
H0: Random Slope Var/Cov = 0	2	12391	38.34	<.0001	MI

MI: P-value based on a mixture of chi-squares.

Model 2C: Adding Cross-Level Interactions

- Level 1:

$$Logit\big(P(Y_{is}=1)\big) = \beta_{0s} + \beta_{1s}(X_{is} - \bar{X}_s) + e_{is}$$

- Level 2:

$$\beta_{0s} = \gamma_{00} + \gamma_{01}\big(\bar{X}_s - \bar{X}_g\big) + U_{0s}$$
$$\beta_{1s} = \gamma_{10} + \gamma_{11}\big(\bar{X}_s - \bar{X}_g\big) + U_{1s}$$

- Combined Model:

$$\begin{aligned} &Logit\big(P(Y_{is}=1)\big) \\ &\quad = \gamma_{00} + \gamma_{10}(X_{is} - \bar{X}_s) \\ &\quad + \gamma_{01}\big(\bar{X}_s - \bar{X}_g\big) + \gamma_{11}(X_{is} - \bar{X}_s)\big(\bar{X}_s - \bar{X}_g\big) + U_{0s} \\ &\quad + U_{1s}(X_{is} - \bar{X}_s) + e_{is} \end{aligned}$$

- Where $e_{is} \sim Logistic(0, \frac{\pi^2}{3})$; $[U_{0s}, U_{1s}] \sim N_2(\mathbf{0}, \boldsymbol{\tau})$

Model 2C: Results

- Model Fit:

Fit Statistics

-2 Log Likelihood	12348.99
AIC (smaller is better)	12362.99
AICC (smaller is better)	12363.00
BIC (smaller is better)	12380.79
CAIC (smaller is better)	12387.79
HQIC (smaller is better)	12370.18

- Estimates:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	schoolID	0.8117	0.1542
UN(2,1)	schoolID	-0.00288	0.002746
UN(2,2)	schoolID	0.000133	0.000050

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-1.2813	0.1013	93	-12.65	<.0001
SMCmath	-0.03552	0.002406	92	-14.76	<.0001
SMmath48	-0.1530	0.01447	12893	-10.57	<.0001
SMCmath*SMmath48	-0.00069	0.000335	12893	-2.06	0.0392

Which Model is Preferred?

- Because we are testing fixed effects, we must form a deviance test by hand
  - Model 2B -2LL: 12,353.04
  - Model 2C -2LL: 12,348.99

- Deviance test: 13,173.52 – 12,391.37 = 782.16
- df = 1
- P-value = 0.044

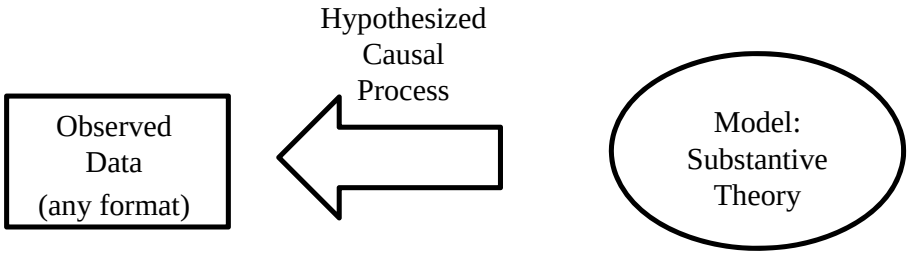
- The p-value is small – we reject the null hypothesis
  - Model 2C is preferred…

GENERALIZED MODELS

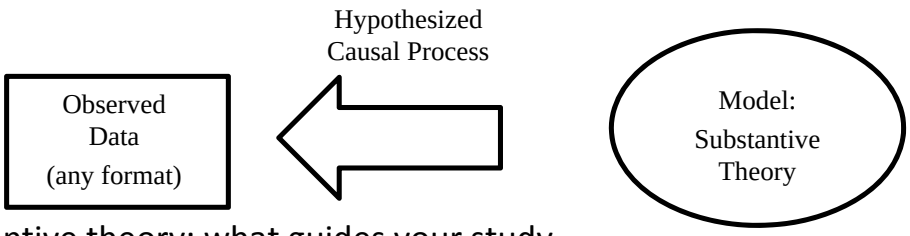
Generalized Models

- Linear models with random effects (AKA latent variables) incorporates a very general set of statistical tools
  - We have only seen tools for use with continuous data that are multivariate normally distributed

- A bigger picture view of the modeling process sees what we know already as one small part



Unpacking the Big Picture



- Substantive theory: what guides your study
  - Examples: one-factor of gambling tendencies; prediction of endogenous variables in path analysis…

- Hypothetical causal process: what the statistical model is testing when estimated

- Observed data: what you collect and evaluate based on your theory
  - Data can take many forms:
    - Continuous variables (e.g., time, blood pressure, height)
    - Categorical variables (e.g., likert-type responses, ordered categories, nominal categories)
    - Combinations of continuous and categorical (e.g., either 0 or some other continuous number)

The Goal of Generalized Models

- Generalized models map the substantive theory onto the space of the observed data
  - Space = type/range/outcomes that are possible
  - Often called sample space in statistics
- The general idea is that the statistical model will not approximate the data well if the assumed distribution is not a good fit to the sample space of the data
- The key to making all of this work is the use of differing statistical distributions

The Basics of Statistical Distributions

- Statistical distributions are functions that describe the probability of a random variable taking certain values
  - In the case of generalized models, we are finding the "right" distribution for our data (the random variables of interest)
- Statistical distributions can be categorized into three classes:
  - Completely continuous
  - Completely categorical (also called discrete)
  - Mixtures of continuous and categorical
- Distributions are defined on a sample space – the range of values random variables can take
  - Univariate normal distribution: $(-\infty, \infty)$ – all real numbers
  - Chi-squared distribution: $[0, \infty)$ – all positive numbers
  - Bernoulli distribution: $\{0,1\}$ – binary digits

More on Distributions

- A statistical distribution has the property that the sum (for categorical) or integral (for continuous) of the distribution equals one across the sample space
  - Subsets of these define the probability of values occurring
- An infinite number of distributions exist – and almost any can be used in generalized models
  - You may have to build your own estimator, though
- More commonly, generalized models allow you to pick from a handful of families of distributions
  - We will stick with what Mplus gives us
- In modern statistical analysis, multiple distributions can be used for different items/variables in an analysis
  - Not every item or variable has to follow one distribution

Link Functions: How Generalized Models Work

- Generalized models work by providing a mapping of the theoretical portion of the model (the right hand side of the equation) to the sample space of the data (the left hand side of the equation)
  - The mapping is done by a feature called a link function
- The link function is a non-linear function that takes the linear model predictors, random/latent terms, and constants and puts them onto the space of the outcome observed variables
- Link functions are typically expressed for the mean of the outcome variable (I will only focus on that)
  - In generalized models, the variance is often a function of the mean

Link Functions in Practice

- The link function expresses the value of the mean of the outcome $E(y_{si}) = \mu_y$ (E stands for expectation)…
- …through a (typically) non-linear function $g(\cdot)$ (when used on the mean; or its inverse $g^{-1}(\cdot)$ when used on the predictors…
- …of the observed predictors (and their regression weights) $\mathbf{X}_s\boldsymbol{\beta}$ …
- …and of the random/latent predictors (and their observed or estimated weights – think factor loadings) $\boldsymbol{Z}_s\boldsymbol{\Gamma}_s$…

$$E(y_{si}) = \mu_y = g^{-1}(\mathbf{X}_s\boldsymbol{\beta} + \boldsymbol{Z}_s\boldsymbol{\Gamma}_s)$$

- The term $\mathbf{X}_s\boldsymbol{\beta} + \boldsymbol{Z}_s\boldsymbol{\Gamma}_s$ is called the **linear predictor**
  - Within the function, the values are linear combinations

CFA in a Generalized Model Context

- A confirmatory factor analysis model is a member of the generalized linear model family
  - The link function is called the identity – it is what it is!
- We knew from before that the expected value of an item from the CFA model is given by:

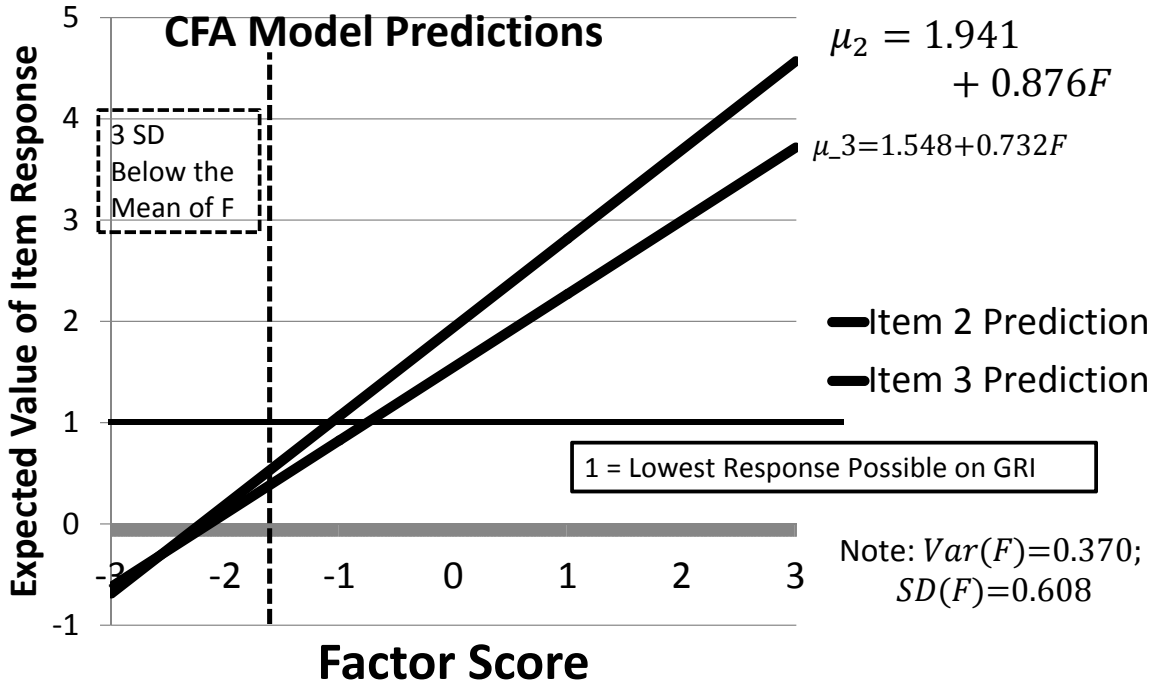
$$E(y_{si}) = \mu_y = g^{-1}\left(\mu_{I_i} + \boldsymbol{\Lambda}_i\mathbf{F}_s\right) = \mu_{I_i} + \boldsymbol{\Lambda}_i\mathbf{F}_s$$

- Here, the inverse link function is the identity

$$g^{-1}(\cdot) = I(\cdot)$$

  - The identity does not alter the predicted values – they can be any real number
  - This matches the sample space of the normal distribution

CFA Model Mean Response Predictions – Using Estimates from 24 Item Analysis



CFA: What About the Variance of the Item?

- The variance of the item (observed outcome data) in a CFA model is given by the estimated unique variance in the model ψ_i^2
- In generalized models, the variance term is often found as a function of the mean (more on that shortly)
- But in this case, we can say:

$$Var(y_{si}) = V(\mu_y) = V(g^{-1}(\mu_{I_i} + \mathbf{\Lambda}_i \mathbf{F}_s)) = \sigma_y^2 = \psi_i^2$$

Putting the Expected Mean/Variance Into a Distribution

- In CFA, we assume our observed data are normally distributed, which means the statistical distribution (conditional on the factor score) for the item is given by:

$$f(y_{si}|\mathbf{F}_s) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{\left(y_{si}-\mu_y\right)^2}{2\sigma_y^2}\right)$$

- Plugging in our model expression for the mean and variance gives:

$$f(y_{si}|\mathbf{F}_s) = \frac{1}{\sqrt{2\pi\psi_i^2}}\exp\left(-\frac{\left(y_{si}-\left(\mu_{I_i}+\boldsymbol{\Lambda}_i\mathbf{F}_s\right)\right)^2}{2\psi_i^2}\right)$$

Where This Is Going…

- In order to explain several key concepts about generalized models, we are going to work through them using our CFA model (identity link; normally distributed data)

- Of importance in comparing GLMM to what we know:
  - Estimation is more complicated (not quite impossible)
  - Evaluation of model fit is more complicated (virtually impossible)

- With CFA (and an identity link), we have a normal distribution assumed for items
  - The normal distribution has two parameters: μ_y, σ_y^2
  - The CFA model makes predictions about these parameters

- In the rest of the generalized models a similar process holds
  - Each statistical distribution has a set of parameters
  - The model makes predictions about the parameters

Generalized Linear Mixed Models

- The overarching name used for linear models with differing types of outcomes (data) and different types of predictors (observed and random/latent variables) is generalized linear mixed models

- This comes from the progression of statistics:
  - Linear models: regression (continuous predictors) w/o random/latent predictors for predicting continuous outcome
  - General linear models: ANOVA (categorical predictors) and regression (continuous predictors) w/o random/latent predictors for predicting continuous outcome
  - Generalized linear models: ANOVA (categorical predictors) and regression (continuous predictors) w/o random/latent predictors for predicting different types of outcomes
  - Generalized linear mixed models: ANOVA (categorical predictors) and regression (continuous predictors) **with** random/latent predictors for predicting different types of outcomes

MARGINAL ML ESTIMATION OF GENERALIZED LINEAR MIXED MODELS

Moving from Marginal (One Item) Distributions to Joint (All Items)

- In order to estimate the model parameters, we need the joint distribution of all of our observed data $f(\mathbf{y}_s)$
  - This joint distribution cannot have any random/latent terms
  - It is just for **all** of the observed data
- At the item level, we have the **conditional** distribution of an item response given our random/latent term (the factor score): $f(y_{si}|\mathbf{F}_s)$
- To get to the joint distribution of the observed data we must go through a series of steps (these are common across GLMMs)
  1. We must first aggregate across all conditional distributions of items to form the joint conditional distribution of all the data $f(\mathbf{y}_s|\mathbf{F}_s)$
     - Still conditional on the random/latent terms
  2. We must then **marginalize** (remove) the random/latent term from the conditional distribution in order to get to the joint distribution of the data $f(\mathbf{y}_s)$

Step #1: The Joint Conditional Distribution

The joint conditional distribution comes from the individual distributions of all of the item responses:

$$f(\mathbf{y}_s|\mathbf{F}_s) = \prod_{i=1}^{I} f(y_{si}|\mathbf{F}_s)$$

This is built from the assumption of item responses being independent given the factor scores (conditional independence) – and gives us the product

Specifically for our data (with a normal distribution) this is:

$$f(\mathbf{y}_s|\mathbf{F}_s) = \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\psi_i^2}} \exp\left(-\frac{\left(y_{si} - \left(\mu_{I_i} + \mathbf{\Lambda}_i\mathbf{F}_s\right)\right)^2}{2\psi_i^2}\right)$$

Pre-Step #2…Mathematical Statistics

- To get to the joint distribution of just the data, we must **marginalize** across the random/latent terms
  - Before we do that, a primer on statistics is in order
- The joint (bivariate) distribution is written as $f(X, Y)$
- The marginal distributions are written as $f(X)$ and $f(Y)$
- Depending on the type of random variable (continuous or discrete) marginal distribution comes from integrating or summing the joint distribution across the sample space of the other variable:

$$f(X) = \int_Y f(X, Y)dY \text{ - continuous}$$
$$f(X) = \sum_Y f(X, Y) \text{ - discrete}$$

Conditional Distributions

- For two random variables X and Y, a conditional distribution is written as: $f(X|Y)$
  - The distribution of X given Y
- The conditional distribution is also equal to the joint distribution divided by the marginal distribution of the conditioning random variable

$$f(X|Y) = \frac{f(X, Y)}{f(Y)}$$

- To get to the marginal (where we need to go) from the conditional (what we have), we have to first get to the joint distribution:

$$f(X) = \int_Y f(X|Y)f(Y)\, dY = \int_Y \frac{f(X, Y)}{f(Y)} f(Y)dY = \int_Y f(X, Y)dY$$

- This is what we will use to get the distribution we are after

Step #2: Marginalizing Across the Random/Latent Terms

The joint marginal distribution of the data $f(\mathbf{y}_s)$ is derived from the same process detailed on the two previous slides:

$$f(\mathbf{y}_s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f(\mathbf{y}_s|\mathbf{F}_s)f(\mathbf{F}_s)\,dF_F \dots dF_2 dF_1$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\left[\prod_{i=1}^{I} f(y_{si}|\mathbf{F}_s)\right] f(\mathbf{F}_s)\,dF_F \dots dF_2 dF_1$$

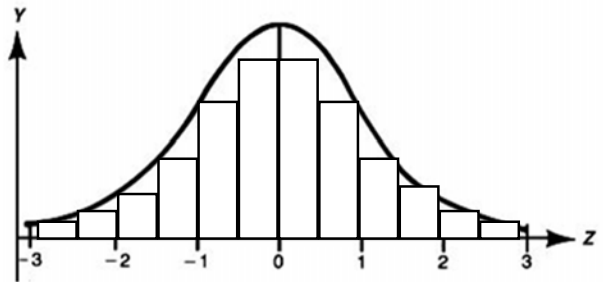
Note: if there is more than one random/latent term, there is more than one integral…one for every random/latent term $F_1, F_2, \dots F_F$

Regardless of the type of item – this marginalization is the same in a GLMM with continuous random/latent terms

We used it in CFA…as we will see

"Marginal" ML Estimation

- How integration works, computationally:
  Divide the distribution into rectangles
  "Gaussian Quadrature" (# rectangles = # "quadrature points")
  - You can either divide the whole distribution into rectangles, or take the most likely section for each person and rectangle that
    - This is "adaptive quadrature" and is computationally more demanding, but gives more accurate results with fewer rectangles



The likelihood of each person's observed data at each value of the random/latent term rectangle is then weighted by that rectangle's probability of being observed (as given by the normal distribution). The weighted likelihoods are then added together across all rectangles.

Distribution of Random/Latent Terms

- Central to the marginalization is the distribution of random/latent terms $f(\mathbf{F}_s)$
  - These are typically assumed to be continuous and normally distributed
- In most GLMMs, these follow a MVN distribution
  - Latent class models and Diagnostic Classification Models use different (categorical) distributions
- The mean of the random/latent terms ($\boldsymbol{\mu}_F$)is usually set to zero, and the covariance matrix ($\boldsymbol{\Phi}$) is estimated:

$$f(\mathbf{F}_s) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Phi}|^{\frac{1}{2}}} \exp\left[-\frac{(\boldsymbol{F}_s^T - \boldsymbol{\mu}_F)^T \boldsymbol{\Phi}^{-1}(\boldsymbol{F}_s^T - \boldsymbol{\mu}_F)}{2}\right]$$

Putting CFA Into The GLMM Context

From previous slides, we found the conditional distribution in CFA to be:

$$f(\mathbf{y}_s|\mathbf{F}_s) = \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\psi_i^2}} \exp\left(-\frac{\left(y_{si} - \left(\mu_{I_i} + \boldsymbol{\Lambda}_i \mathbf{F}_s\right)\right)^2}{2\psi_i^2}\right)$$

We also found the distribution of the latent factors to be:

$$f(\mathbf{F}_s) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Phi}|^{\frac{1}{2}}} \exp\left[-\frac{(\boldsymbol{F}_s^T - \boldsymbol{\mu}_F)^T \boldsymbol{\Phi}^{-1}(\boldsymbol{F}_s^T - \boldsymbol{\mu}_F)}{2}\right]$$

Putting CFA Into The GLMM Context

Putting these together, we get:

$$f(\mathbf{y}_s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(\mathbf{y}_s|\mathbf{F}_s)f(\mathbf{F}_s)\,dF_F \ldots dF_2 dF_1$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\left[\prod_{i=1}^{I} f(y_{si}|\mathbf{F}_s)\right] f(\mathbf{F}_s)\,dF_F \ldots dF_2 dF_1$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\left[\prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\psi_i^2}}\exp\left(-\frac{\left(y_{si}-\left(\mu_{I_i}+\boldsymbol{\Lambda}_i\mathbf{F}_s\right)\right)^2}{2\psi_i^2}\right)\right] x$$

$$\frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Phi}|^{\frac{1}{2}}}\exp\left[-\frac{(\boldsymbol{F}_s^T-\boldsymbol{\mu}_F)^T\boldsymbol{\Phi}^{-1}(\boldsymbol{F}_s^T-\boldsymbol{\mu}_F)}{2}\right] dF_F \ldots dF_2 dF_1$$

OMFG!

CFA Relies on MVN Properties to Simplify

- The monstrous equation from the last slide has an easier version – all due to properties of MVN distributions
  - Conditional distributions of MVN are also MVN
  - Marginal distributions of MVNs are also MVN

- Therefore, we can show that for CFA (under identification where the factor mean is zero), the last slide becomes:

$$f(\mathbf{y}_s) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T+\boldsymbol{\Psi}|^{\frac{1}{2}}}\exp\left[-\frac{(\mathbf{y}_s^T-\boldsymbol{\mu}_I)^T(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T+\boldsymbol{\Psi})^{-1}(\mathbf{y}_s^T-\boldsymbol{\mu}_I)}{2}\right]$$

What All This Means

- The integrals in the non-specific GLMM are difficult to estimate computationally
  - They take a long time – and get approximated
  - CFA doesn't have them because of the MVN distribution
- Model fit is based on the joint distribution of the data $f(\mathbf{y}_s)$, across all subjects s, or $f(\mathbf{Y})$
  - In general, this is difficult to impossible to figure out for differing distributions in the GLMM
  - CFA doesn't have this problem as the joint distribution is MVN
- Therefore, two fundamental aspects of CFA don't map well onto GLMMs
  - Easy estimation
  - Relatively easy model fit determination

WRAPPING UP

Generalized Linear Mixed Models

- Generalized Linear Mixed Models are flexible models that allow for the same multilevel modeling process to happen for a variety of types of data

- In fact, we could have used these models as IRT models
  - Explanatory IRT models are very similar

- This lecture demonstrated a little bit of the modeling process for such models
  - Much more to go