LMMs in Practice

Jin Karlen

April 22, 2010

Overview

- **Problem and Motivation** (2 min.)
- **Model-Building Strategies** (3 min.)
- **Two Examples**
  1. **Two-Level Models for Clustered Data :** The Rat Pup Example (15 min.)
  2. **Random Coefficient Models for Longitudinal Data:** The Autism Example (25 min.)
     - ➢ Structures of Analyzing
     - ➢ The Autism Example in R

Problem and Motivation

- **What is important in an application of LMM ?**
  - ✓ **Dependent variable**
  - ✓ **Covariances: fixed-effect parameters and random-effect parameters**
  - ✓ **The relationships between a continuous dependent variable and various predictor variables**
- **What kind of data sets are they?**
  - ✓ **Clustered, longitudinal, or repeated-measures**
- **How can we analyze those data?**
- **How can we build a suitable model?**
- **How can we know, if it is a good model?**
- **…….**

Model-Building Strategies

✓The Top-Down Strategy

1. Start with a well-specified mean structure for the model
2. Select a structure for the random effects in the model
3. Select a covariance structure for the residuals in the model
4. Reduce the model

✓The Step-Up Strategy

✓….

The Rat Pup Study

- **Litter (Level 2) Variables**

LITTER = Litter ID number

TREATMENT = Dose level of the experimental compound assigned to the litter(high, low, control)

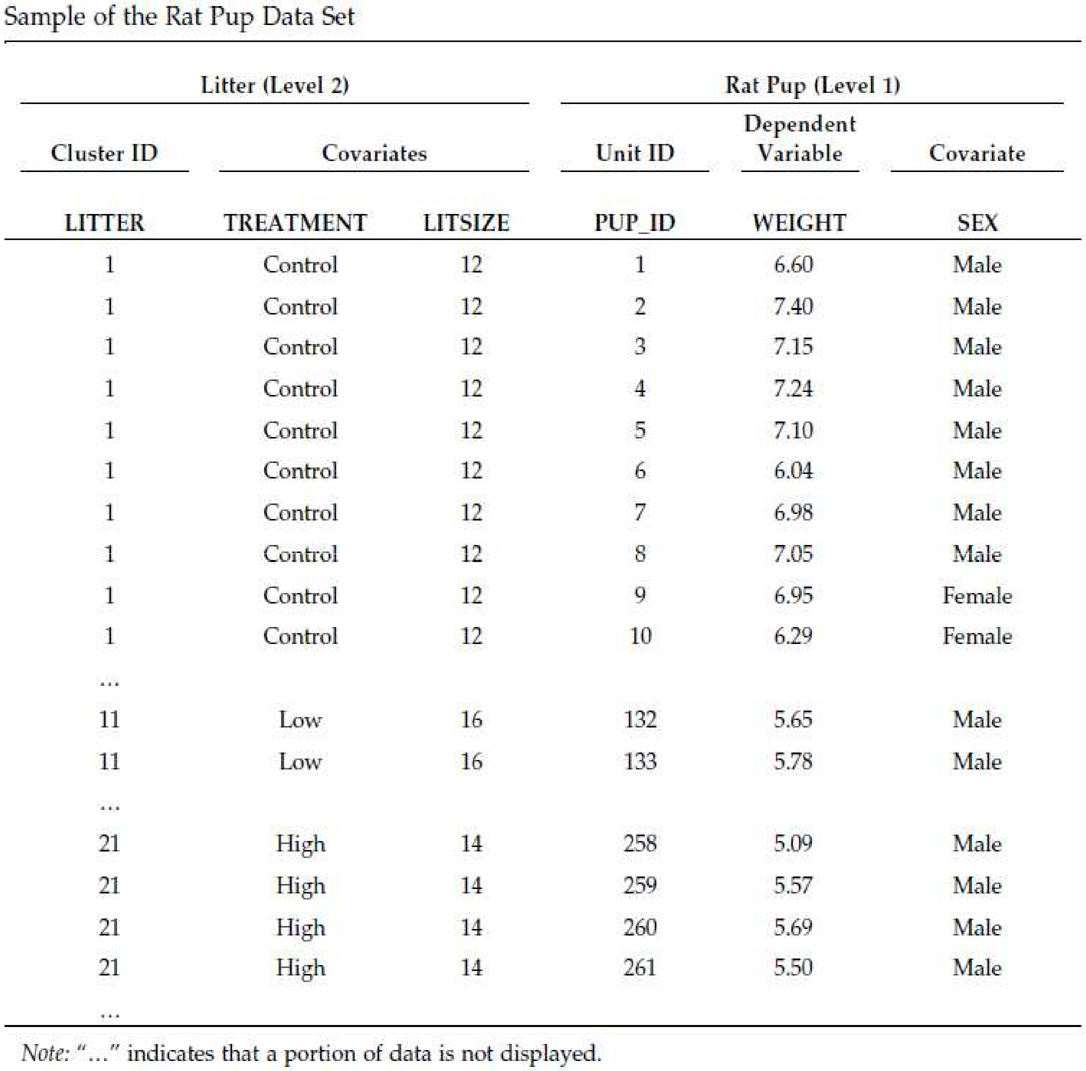
LITSIZE = Litter size (i.e., number of pups per litter)

- **Rat Pup (Level 1) Variables**

PUP_ID = Unique identifier for each rat pup

WEIGHT = Birth weight of the rat pup (the dependent variable)

SEX = Sex of the rat pup (male, female)

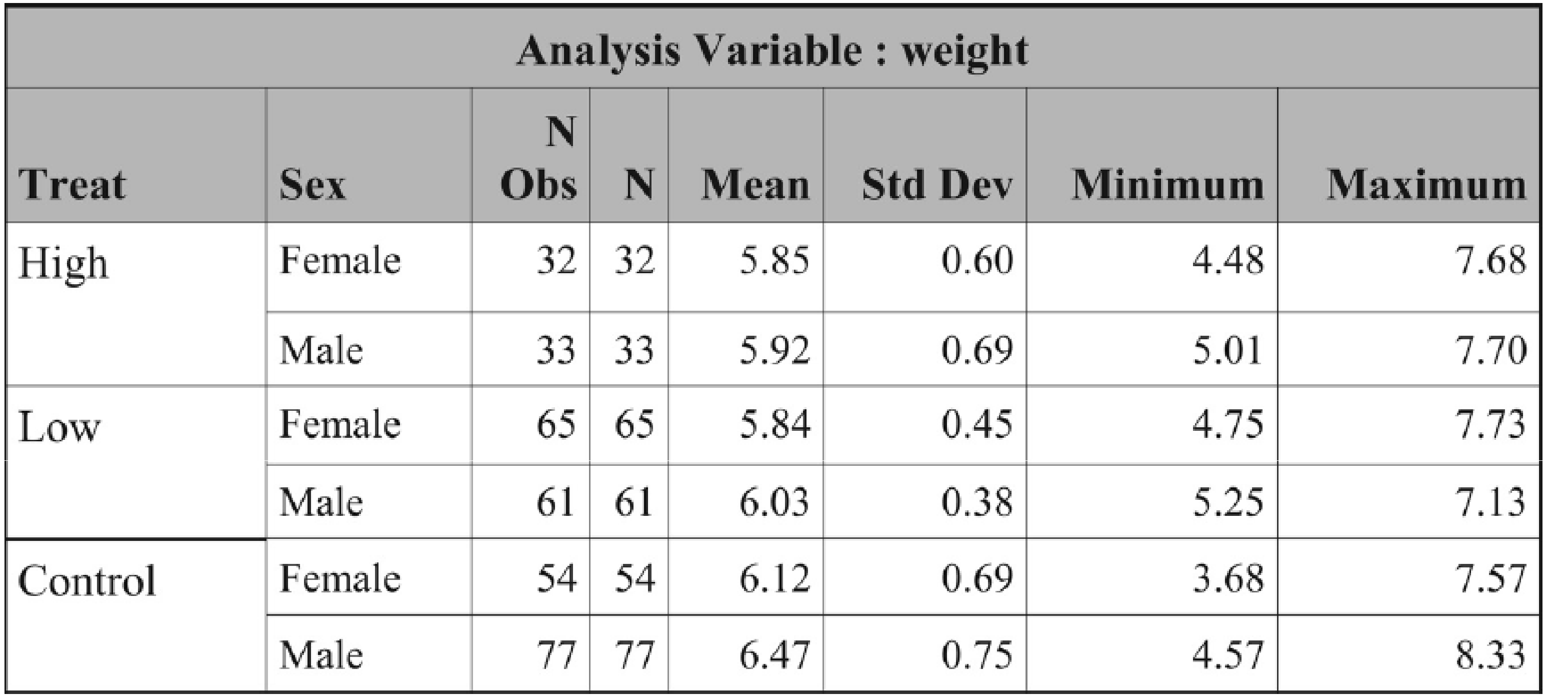


Sample of the Rat Pup Data Set

Litter (Level 2)			Rat Pup (Level 1)		
Cluster ID	Covariates		Unit ID	Dependent Variable	Covariate
LITTER	TREATMENT	LITSIZE	PUP_ID	WEIGHT	SEX
1	Control	12	1	6.60	Male
1	Control	12	2	7.40	Male
1	Control	12	3	7.15	Male
1	Control	12	4	7.24	Male
1	Control	12	5	7.10	Male
1	Control	12	6	6.04	Male
1	Control	12	7	6.98	Male
1	Control	12	8	7.05	Male
1	Control	12	9	6.95	Female
1	Control	12	10	6.29	Female
…					
11	Low	16	132	5.65	Male
11	Low	16	133	5.78	Male
…					
21	High	14	258	5.09	Male
21	High	14	259	5.57	Male
21	High	14	260	5.69	Male
21	High	14	261	5.50	Male
…					

Note: "…" indicates that a portion of data is not displayed.

Data Summary



Analysis Variable : weight

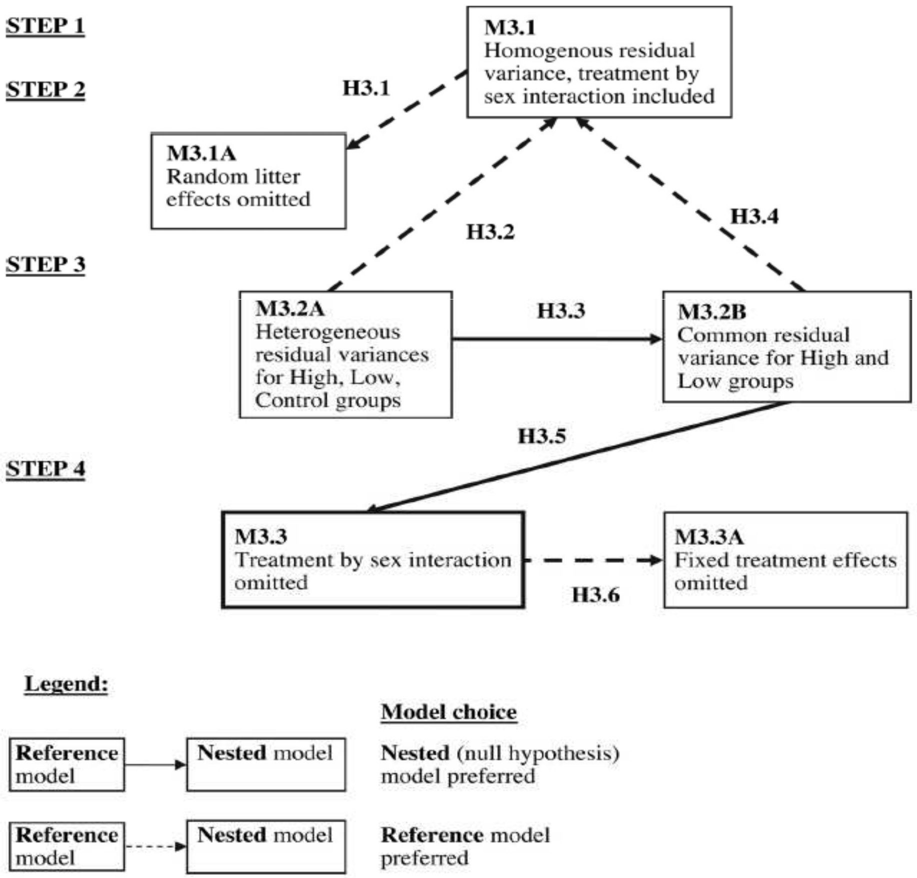
Treat	Sex	N Obs	N	Mean	Std Dev	Minimum	Maximum
High	Female	32	32	5.85	0.60	4.48	7.68
	Male	33	33	5.92	0.69	5.01	7.70
Low	Female	65	65	5.84	0.45	4.75	7.73
	Male	61	61	6.03	0.38	5.25	7.13
Control	Female	54	54	6.12	0.69	3.68	7.57
	Male	77	77	6.47	0.75	4.57	8.33

LITTER (number of litters) = 27
LITSIZE (number of rat pups per litter) = 2~18
The number of pups = 322
Female (rat pups **) =** 151
Male (rat pups **) =** 171
WEIGHT (Birth weight of the rat pup) = 3.68 ~ 8.33

Model Specification

$$\left.\begin{aligned}\text{WEIGHT}_{ij} &= \beta_0 + \beta_1 \times \text{TREAT1}_j + \beta_2 \times \text{TREAT2}_j + \beta_3 \times \text{SEX1}_{ij} \\ &+ \beta_4 \times \text{LITSIZE}_j + \beta_5 \times \text{TREAT1}_j \times \text{SEX1}_{ij} + \beta_6 \times \text{TREAT2}_j \times \text{SEX1}_{ij}\end{aligned}\right\}\text{fixed}$$

$$+ u_j + \varepsilon_{ij} \} \text{ random}$$



STEP 1

STEP 2

STEP 3

STEP 4

M3.1 Homogenous residual variance, treatment by sex interaction included

M3.1A Random litter effects omitted

M3.2A Heterogeneous residual variances for High, Low, Control groups

M3.2B Common residual variance for High and Low groups

M3.3 Treatment by sex interaction omitted

M3.3A Fixed treatment effects omitted

H3.1

H3.2

H3.3

H3.4

H3.5

H3.6

Legend:

Model choice

Reference model → **Nested** model: **Nested** (null hypothesis) model preferred

Reference model - - → **Nested** model: **Reference** model preferred

FIGURE 3.3
Model selection and related hypotheses for the analysis of the Rat Pup data.

Analysis Steps

- **Step 1: Fit a model with a "loaded" mean structure (Model 3.1).**

  Model 3.1 includes treatment, sex, litter size, interaction between treatment and sex, random effect associated with the intercept for each litter and a residual (i.i.d.) associated with each birth weight observation.

- **Step 2: Select a structure for the random effects (Model 3.1 vs. Model 3.1A).**

  Model 3.1 A : by omitted the random litter effects from Model 3.1 (Hypothesis 3.1).

- **Step 3: Select a covariance structure for the residuals (Model 3.1, Model 3.2A, or Model 3.2B).**

  Model 3.1 : homogeneous residual for all treatment groups

  Model 3.2A: heterogeneous residual for for each level of treatment (high, low, and control).

  Model 3.2B: a common residual variance for the high and low treatment groups, and a different residual variance for the control group.

- **Step 4: Reduce the model by removing nonsignificant fixed effects, test the main effects associated with treatment, and assess model diagnostics.**

  Decide whether to keep the treatment by sex interaction in Model 3.2B (Model 3.2B vs. Model 3.3).

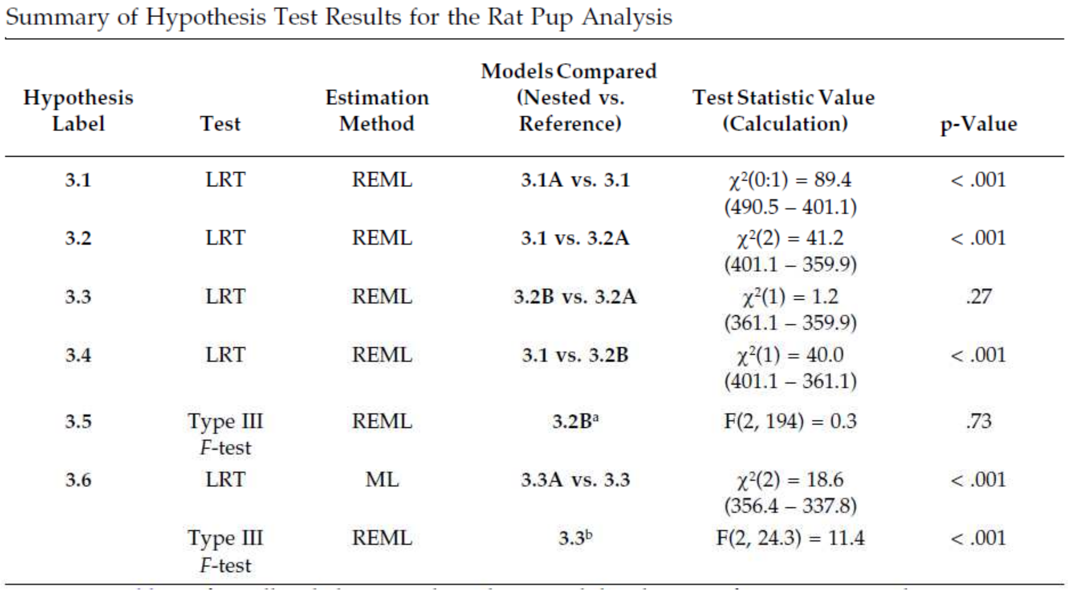
  Test the significance of the treatment effects in our final model, Model 3.3 (Model 3.3 vs. Model 3.3A).

  Assess the assumptions for Model 3.3.

Hypothesis Tests

- Hypothesis 3.1: The random effects, uj, associated with the litter-specific intercepts can be omitted from Model 3.1.
- Hypothesis 3.2: The variance of the residuals is the same (homogeneous) for the three treatment groups (high, low, and control).
- Hypothesis 3.3: The residual variances for the high and low treatment groups are equal.
- Hypothesis 3.4: The residual variance for the combined high/low treatment group is equal to the residual variance for the control group.
- Hypothesis 3.5: The fixed effects associated with the treatment by sex interaction are equal to zero in Model 3.2B.

Hypothesis 3.6: The fixed effects associated with treatment are equal to zero in Model 3.3.



Summary of Hypothesis Test Results for the Rat Pup Analysis

Hypothesis Label	Test	Estimation Method	Models Compared (Nested vs. Reference)	Test Statistic Value (Calculation)	p-Value
3.1	LRT	REML	3.1A vs. 3.1	$\chi^2(0{:}1) = 89.4$ (490.5 – 401.1)	< .001
3.2	LRT	REML	3.1 vs. 3.2A	$\chi^2(2) = 41.2$ (401.1 – 359.9)	< .001
3.3	LRT	REML	3.2B vs. 3.2A	$\chi^2(1) = 1.2$ (361.1 – 359.9)	.27
3.4	LRT	REML	3.1 vs. 3.2B	$\chi^2(1) = 40.0$ (401.1 – 361.1)	< .001
3.5	Type III *F*-test	REML	3.2B[a]	F(2, 194) = 0.3	.73
3.6	LRT	ML	3.3A vs. 3.3	$\chi^2(2) = 18.6$ (356.4 – 337.8)	< .001
	Type III *F*-test	REML	3.3[b]	F(2, 24.3) = 11.4	< .001

$$\left.\begin{array}{l}\text{WEIGHT}_{ij} = \beta_0 + \beta_1 \times \text{TREAT1}_j + \beta_2 \times \text{TREAT2}_j + \beta_3 \times \text{SEX1}_{ij} \\ + \beta_4 \times \text{LITSIZE}_j + \beta_5 \times \cancel{\text{TREAT1}_j \times \text{SEX1}_{ij}} + \beta_6 \times \cancel{\text{TREAT2}_j \times \text{SEX1}_{ij}}\end{array}\right\} \text{fixed}$$

$$+ u_j + \varepsilon_{ij} \} \text{ random}$$

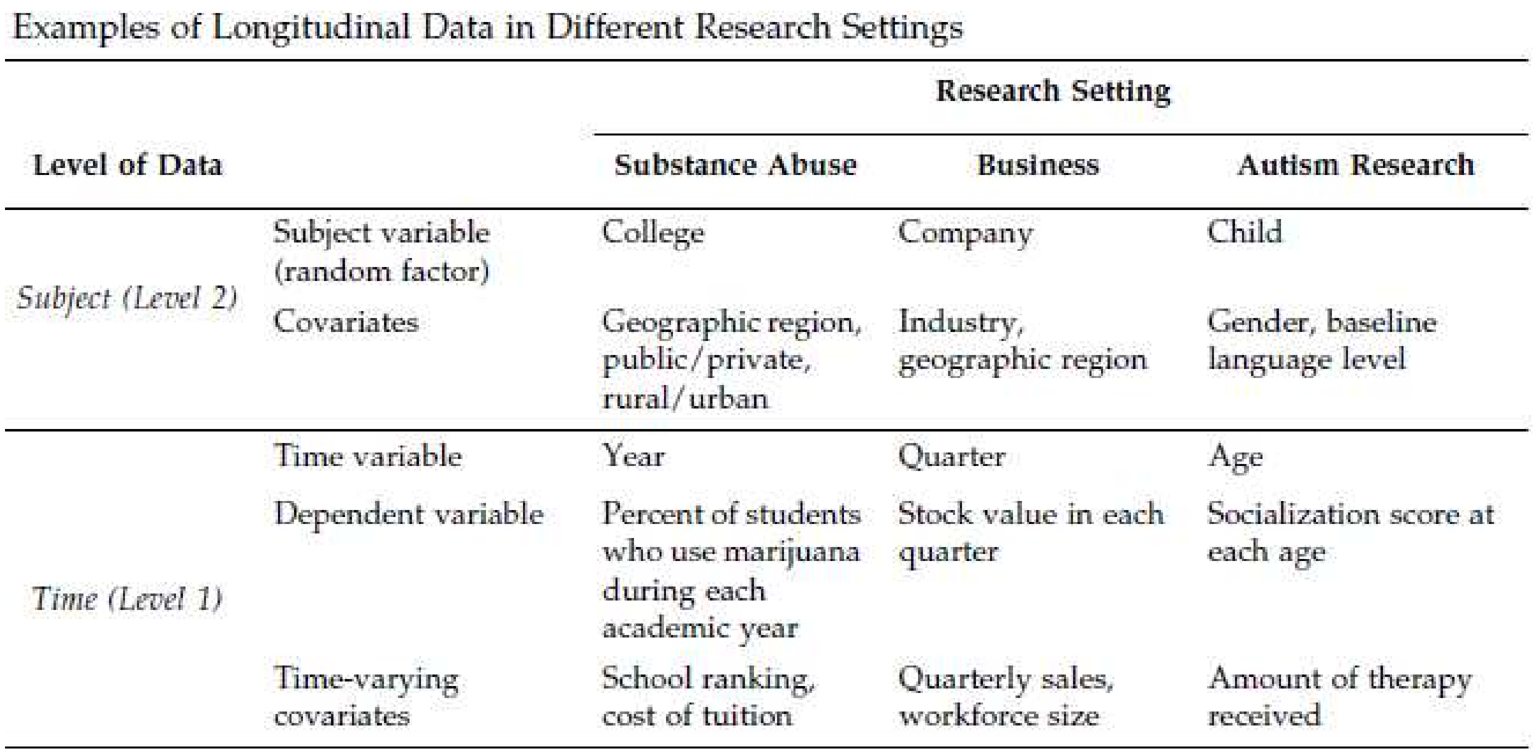
Selected Models Considered in the Analysis of the Rat Pup Data

		Term/Variable	General Notation	HLM Notation	Model			
					3.1	3.2A[a]	3.2B[a]	3.3[a]
Fixed effects		**Intercept**	β_0	γ_{00}	√	√	√	√
		TREAT1 (High vs. control)	β_1	γ_{02}	√	√	√	√
		TREAT2 (Low vs. control)	β_2	γ_{03}	√	√	√	√
		SEX1 (Female vs. male)	β_3	γ_{10}	√	√	√	√
		LITSIZE	β_4	γ_{01}	√	√	√	√
		TREAT1 × SEX1	β_5	γ_{11}	√	√	√	
		TREAT2 × SEX1	β_6	γ_{12}	√	√	√	
Random effects	**Litter (*j*)**	**Intercept**	u_j	u_{0j}	√	√	√	√
Residuals	**Rat pup (pup *i* in litter *j*)**		ε_{ij}	r_{ij}	√	√	√	√
Covariance parameters (θ_D) for *D* matrix	**Litter level**	**Variance of intercepts**	σ^2_{litter}	τ	√	√	√	√
Covariance parameters (θ_R) for R_i matrix	**Rat pup level**	**Variances of residuals**	$\sigma^2_{high}\sigma^2_{low}$ $\sigma^2_{control}$	σ^2	σ^2	σ^2_{high} σ^2_{low} $\sigma^2_{control}$	$\sigma^2_{high/low},$ $\sigma^2_{control}$	$\sigma^2_{high/low},$ $\sigma^2_{control}$

[a] Models 3.2A, 3.2B, and 3.3 (with heterogeneous residual variances) can only be fit using the procedures in SAS and R.

Random Coefficient Models for Longitudinal Data

- **Definition of Longitudinal Data**: data sets in which the dependent variable is measured at several points in time for each unit of analysis.



Examples of Longitudinal Data in Different Research Settings

		Research Setting		
Level of Data		**Substance Abuse**	**Business**	**Autism Research**
Subject (Level 2)	Subject variable (random factor)	College	Company	Child
	Covariates	Geographic region, public/private, rural/urban	Industry, geographic region	Gender, baseline language level
Time (Level 1)	Time variable	Year	Quarter	Age
	Dependent variable	Percent of students who use marijuana during each academic year	Stock value in each quarter	Socialization score at each age
	Time-varying covariates	School ranking, cost of tuition	Quarterly sales, workforce size	Amount of therapy received

The Autism Example

- **Subject (Level 2) Variables**

CHILDID = Unique child identifier

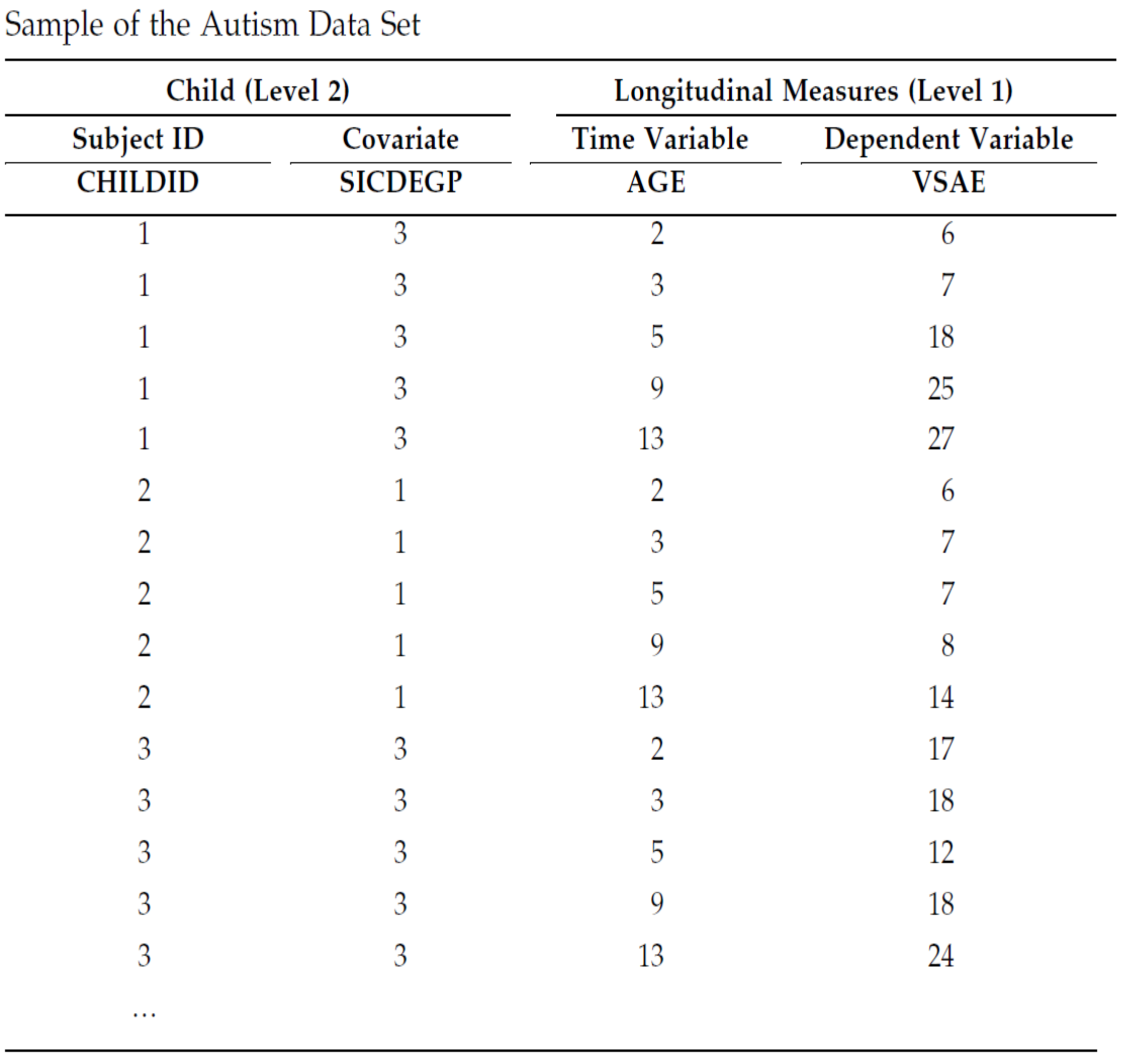
SICDEGP = Sequenced Inventory of Communication Development Expressive

Group: categorized expressive language score at age 2 years (1 = low, 2 =medium, 3 = high)

- **Time-Varying (Level 1) Variables**

AGE = Age in years (2, 3, 5, 9, 13); the time variable

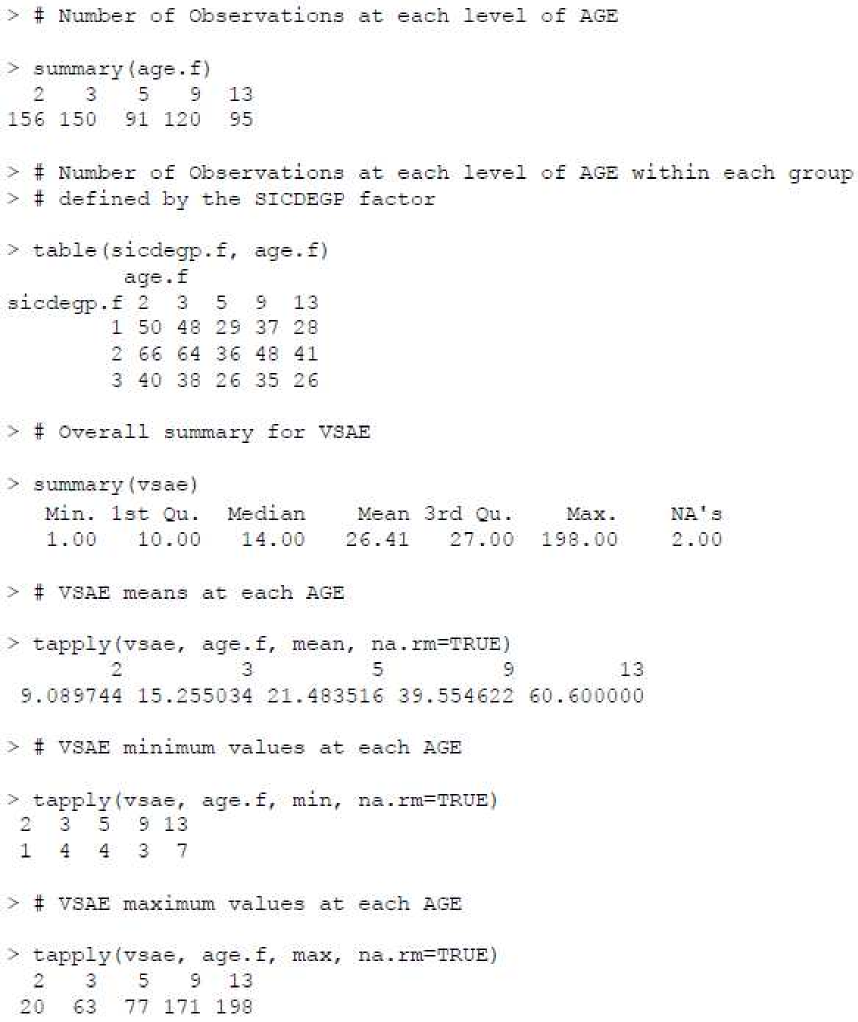
VSAE = Vineland Socialization Age Equivalent: parent-reported socialization, the dependent variable, measured at each age



Sample of the Autism Data Set

Child (Level 2)		Longitudinal Measures (Level 1)	
Subject ID	Covariate	Time Variable	Dependent Variable
CHILDID	SICDEGP	AGE	VSAE
1	3	2	6
1	3	3	7
1	3	5	18
1	3	9	25
1	3	13	27
2	1	2	6
2	1	3	7
2	1	5	7
2	1	9	8
2	1	13	14
3	3	2	17
3	3	3	18
3	3	5	12
3	3	9	18
3	3	13	24
…			

Data Summary



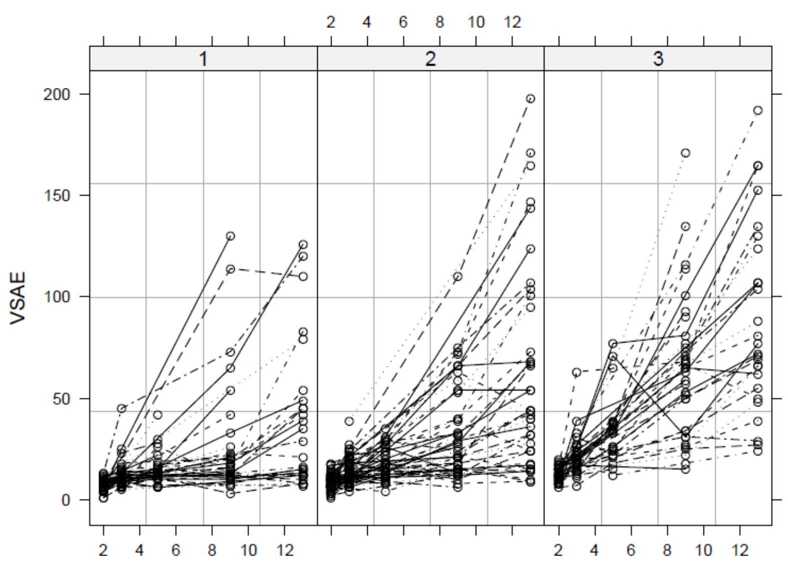
```
> # Number of Observations at each level of AGE

> summary(age.f)
  2   3   5   9  13
156 150  91 120  95

> # Number of Observations at each level of AGE within each group
> # defined by the SICDEGP factor

> table(sicdegp.f, age.f)
         age.f
sicdegp.f 2  3  5  9  13
        1 50 48 29 37 28
        2 66 64 36 48 41
        3 40 38 26 35 26

> # Overall summary for VSAE

> summary(vsae)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1.00   10.00   14.00   26.41   27.00  198.00    2.00

> # VSAE means at each AGE

> tapply(vsae, age.f, mean, na.rm=TRUE)
        2         3         5         9        13
 9.089744 15.255034 21.483516 39.554622 60.600000

> # VSAE minimum values at each AGE

> tapply(vsae, age.f, min, na.rm=TRUE)
 2  3  5  9 13
 1  4  4  3  7

> # VSAE maximum values at each AGE

> tapply(vsae, age.f, max, na.rm=TRUE)
  2   3   5   9  13
 20  63  77 171 198
```

- We begin by reading the comma-separated raw data file (autism.csv) into R functions

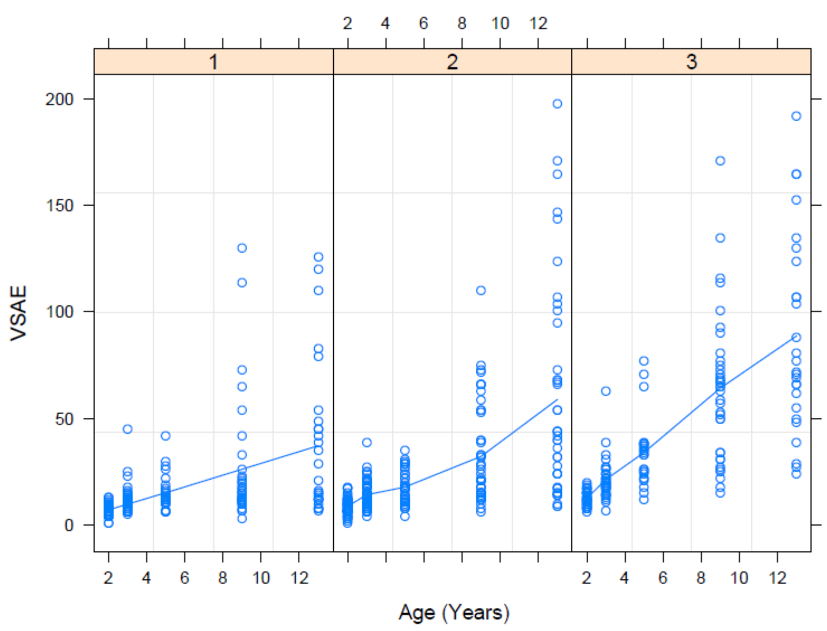
- Next, we apply the factor() function to the numeric variables SICDEGP and AGE to create categorical versions of these variables (SICDEGP.F and AGE.F),

- Add the new variables to the data frame object. After creating these factors, we request descriptive statistics for both the continuous and factor variables included in the analysis using the summary() function

- We next generate graphs that show the observed VSAE scores as a function of age for each child within levels of SICDEGP (Figure 6.1) and the mean VSAE profiles by SICDEGP (Figure 6.2).



Individual Data by SICD Group



Mean Profiles by SICD Group

Result of Data Summary

- The plots of the observed VSAE values for individual children in Figure 6.1 show substantial variation from child to child within each level of SICD group. the VSAE scores of some children tend to increase as the children get older, for other children remain relatively constant. we do not see much variability in the initial values of VSAE at age 2 years for any of the levels of the SICD group. Overall.
- The mean profiles displayed in Figure 6.2 show that mean VSAE scores generally increase with age. There may also be a quadratic trend in VSAE scores, especially in SICD group two. This suggests that a model to predict VSAE should include both linear and quadratic fixed effects of age, and possibly interactions between the linear and quadratic effects of age and SICD group.

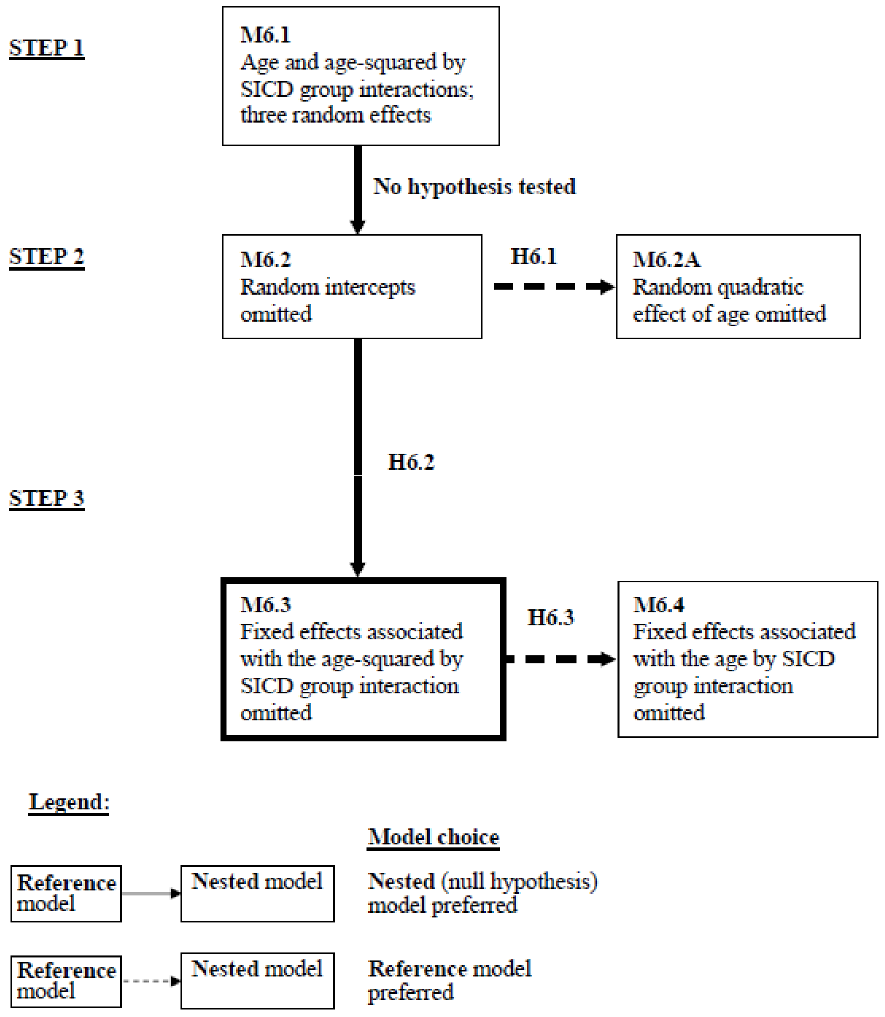
General Model Specification

$$\left.\begin{array}{l} \text{VSAE}_{ti} = \beta_0 + \beta_1 \times \text{AGE\_2}_{ti} + \beta_2 \times \text{AGE\_2SQ}_{ti} + \beta_3 \times \text{SICDEGP1}_i \\ + \beta_4 \times \text{SICDEGP2}_i + \beta_5 \times \text{AGE\_2}_{ti} \times \text{SICDEGP1}_i \\ + \beta_6 \times \text{AGE\_2}_{ti} \times \text{SICDEGP2}_i + \beta_7 \times \text{AGE\_2SQ}_{ti} \times \text{SICDEGP1}_i \\ + \beta_8 \times \text{AGE\_2SQ}_{ti} \times \text{SICDEGP2}_i + \end{array}\right\} \;\textbf{fixed} \qquad (6.1)$$

$$u_{0i} + u_{1i} \times \text{AGE\_2}_{ti} + u_{2i} \times \text{AGE\_2SQ}_{ti} + \varepsilon_{ti} \} \quad \textbf{random}$$

- VSAE_{ti} (Vineland Socialization Age Equivalent*): on child i, at the t-th* visit (*t = 1, 2, 3, 4, 5, corresponding to ages 2, 3, 5, 9 and 13)*
- SICDEGP1 and SICDEGP2: the first two levels of the SICD group, SICDEGP = 3 as the “reference category.”
- AGE_2. SICDEGP1 and AGE_2 . SICDEGP2: interaction between age and SICD group
- AGE_2 SQ . SICDEGP1 and AGE_2SQ . SICDEGP2: interaction between age-squared and SICD group
- β0 - β8 :the fixed effects associated with the intercept,the covariates, and the interaction terms in the model.
- *u0i*, *u*1*i*, *u*2*i* the random effects associated with the child-specific intercept, linear effect of age, and quadratic effect of age for child *i*.
- ε*ti* in Equation 6.1 represents the residual associated with the observation at time *t* on child *i*. ε*ti* ~ *N(0,σ2)*

Overview of the Autism Data Analysis


STEP 1
M6.1
Age and age-squared by SICD group interactions; three random effects
No hypothesis tested
STEP 2
M6.2
Random intercepts omitted
H6.1
M6.2A
Random quadratic effect of age omitted
H6.2
STEP 3
M6.3
Fixed effects associated with the age-squared by SICD group interaction omitted
H6.3
M6.4
Fixed effects associated with the age by SICD group interaction omitted
Legend:
Model choice
Reference model
Nested model
Nested (null hypothesis) model preferred
Reference model
Nested model
Reference model preferred

- **Step 1: Fit a model with a "loaded" mean structure (Model 6.1).**

- **Step 2: Select a structure for the random effects (Model 6.2 vs. Model 6.2A).**

  Fit a model without the random child-specific intercepts (Model 6.2), and test, whether to keep the remaining random effects in the model.

- **Step 3: Reduce the model by removing nonsignificant fixed effects (Model 6.2 vs.**

  Model 6.3), and check model diagnostics.

Hypothesis Tests

- **Hypothesis 6.1: The random effects associated with the quadratic effect of AGE can be omitted from Model 6.2.**

- **Hypothesis 6.2: The fixed effects associated with the AGE-squared . SICDEGP interaction are equal to zero in Model 6.2.**

- **Hypothesis 6.3: The fixed effects associated with the AGE . SICDEGP interaction are equal to zero in Model 6.3.**

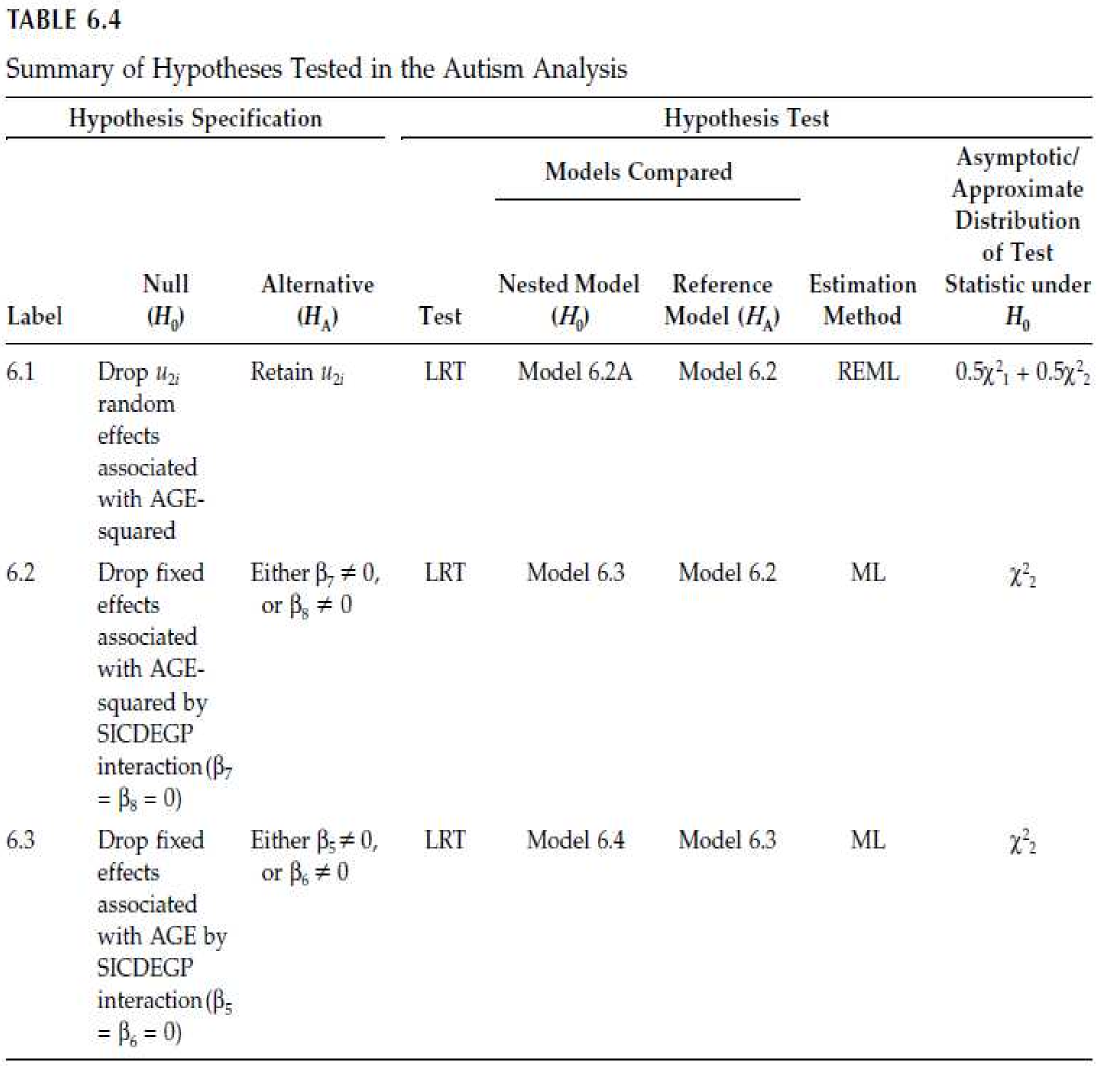
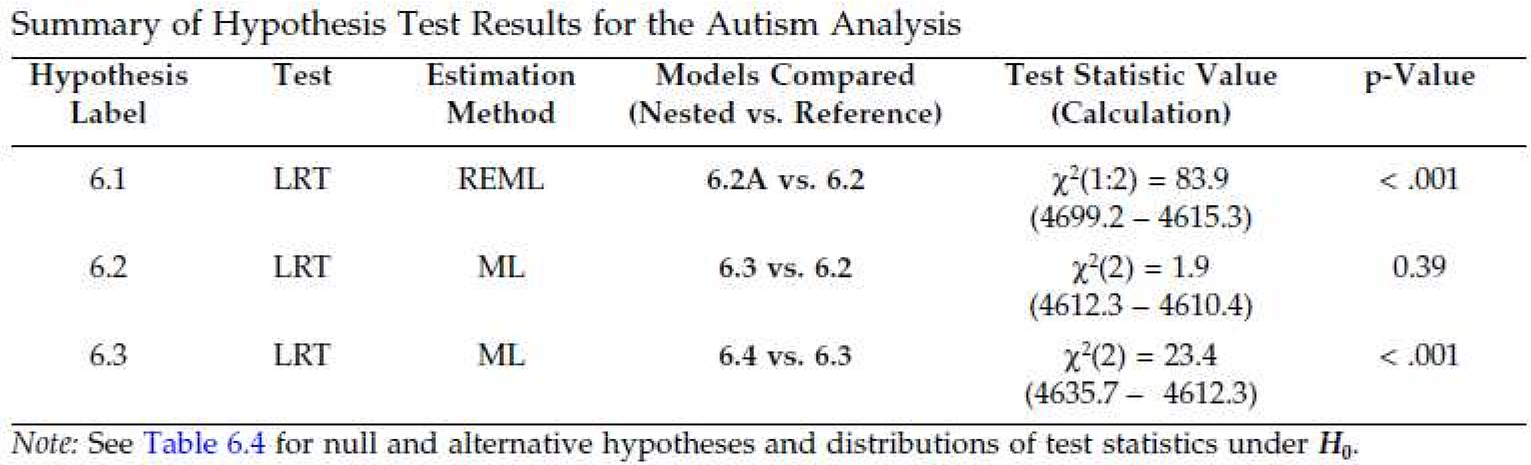


TABLE 6.4

Summary of Hypotheses Tested in the Autism Analysis

Hypothesis Specification			Hypothesis Test				
				Models Compared			Asymptotic/ Approximate Distribution of Test Statistic under H_0
Label	Null (H_0)	Alternative (H_A)	Test	Nested Model (H_0)	Reference Model (H_A)	Estimation Method	
6.1	Drop u_{2i} random effects associated with AGE-squared	Retain u_{2i}	LRT	Model 6.2A	Model 6.2	REML	$0.5\chi^2_1 + 0.5\chi^2_2$
6.2	Drop fixed effects associated with AGE-squared by SICDEGP interaction ($\beta_7 = \beta_8 = 0$)	Either $\beta_7 \neq 0$, or $\beta_8 \neq 0$	LRT	Model 6.3	Model 6.2	ML	χ^2_2
6.3	Drop fixed effects associated with AGE by SICDEGP interaction ($\beta_5 = \beta_6 = 0$)	Either $\beta_5 \neq 0$, or $\beta_6 \neq 0$	LRT	Model 6.4	Model 6.3	ML	χ^2_2

Results of Hypothesis Tests



Summary of Hypothesis Test Results for the Autism Analysis

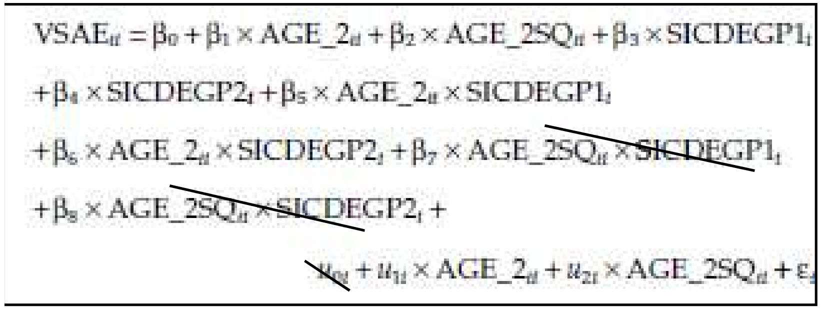
Hypothesis Label	Test	Estimation Method	Models Compared (Nested vs. Reference)	Test Statistic Value (Calculation)	p-Value
6.1	LRT	REML	6.2A vs. 6.2	$\chi^2(1:2) = 83.9$ (4699.2 – 4615.3)	< .001
6.2	LRT	ML	6.3 vs. 6.2	$\chi^2(2) = 1.9$ (4612.3 – 4610.4)	0.39
6.3	LRT	ML	6.4 vs. 6.3	$\chi^2(2) = 23.4$ (4635.7 – 4612.3)	< .001

Note: See Table 6.4 for null and alternative hypotheses and distributions of test statistics under H_0.

- **Hypothesis 6.1: The child-specific quadratic random effects of age can be omitted from Model 6.2.**
- **Hypothesis 6.2: The age-squared by SICD group interaction effects can be dropped from Model 6.2 (β7 = β8 = 0).**
- **Hypothesis 6.3: The age by SICD group interaction effects can be dropped from Model 6.3 (β5 = β6 = 0).**

Diagnostics for the Final Model

- **Residual Diagnostics**
- **Diagnostics for the Random Effects**
- **Observed and Predicted Values**



$$\begin{aligned} \text{VSAE}_{ti} &= \beta_0 + \beta_1 \times \text{AGE\_2}_{ti} + \beta_2 \times \text{AGE\_2SQ}_{ti} + \beta_3 \times \text{SICDEGP1}_i \\ &+ \beta_4 \times \text{SICDEGP2}_i + \beta_5 \times \text{AGE\_2}_{ti} \times \text{SICDEGP1}_i \\ &+ \beta_6 \times \text{AGE\_2}_{ti} \times \text{SICDEGP2}_i + \beta_7 \times \text{AGE\_2SQ}_{ti} \cancel{\times \text{SICDEGP1}_i} \\ &+ \beta_8 \times \text{AGE\_}\cancel{\text{2SQ}_{ti} \times \text{SICDEGP2}_i} + \\ &\qquad \cancel{u_{0i}} + u_{1i} \times \text{AGE\_2}_{ti} + u_{2i} \times \text{AGE\_2SQ}_{ti} + \varepsilon \end{aligned}$$

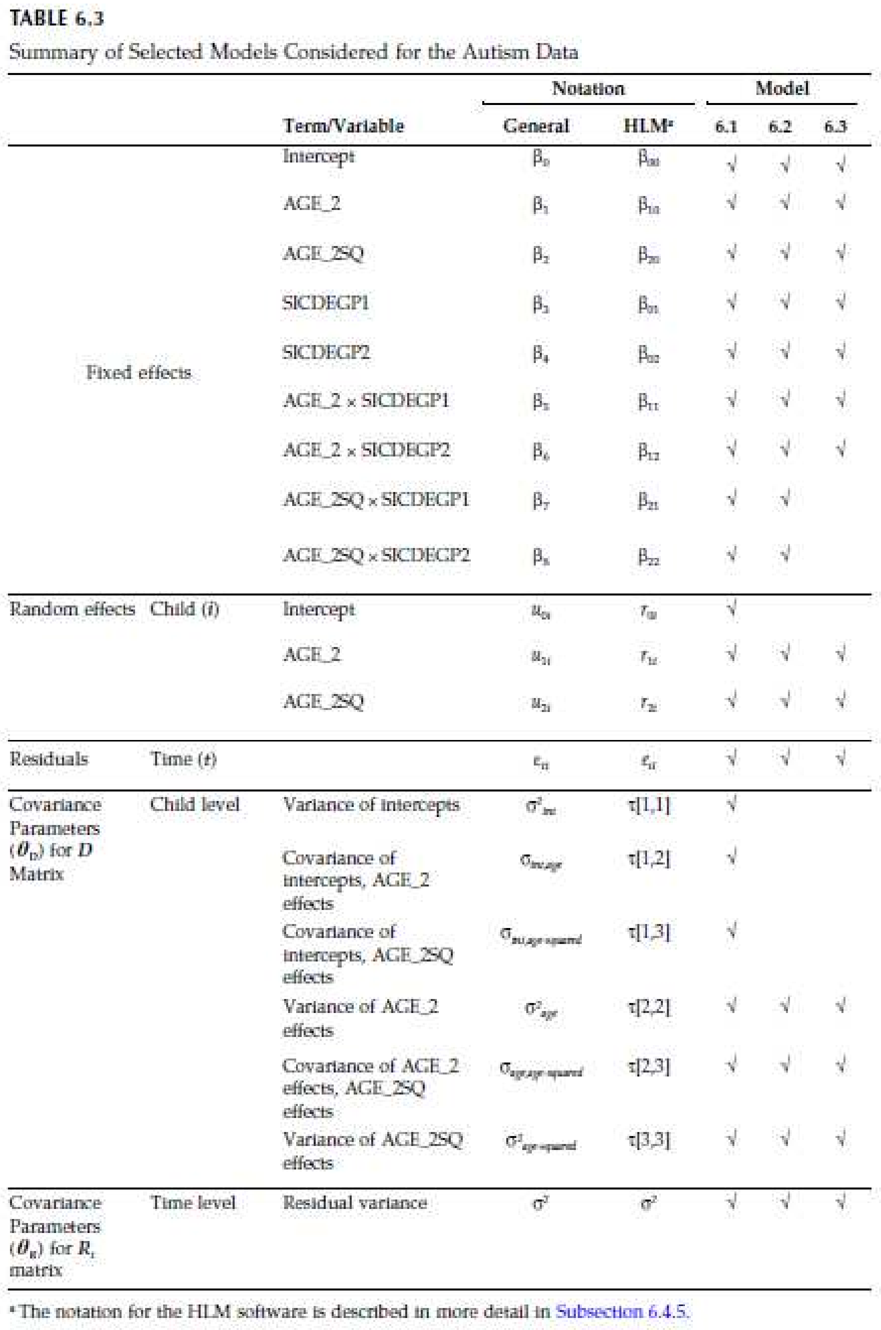


TABLE 6.3

Summary of Selected Models Considered for the Autism Data

			Notation		Model		
		Term/Variable	General	HLM[a]	6.1	6.2	6.3
Fixed effects		Intercept	β_0	β_{00}	√	√	√
		AGE_2	β_1	β_{10}	√	√	√
		AGE_2SQ	β_2	β_{20}	√	√	√
		SICDEGP1	β_3	β_{01}	√	√	√
		SICDEGP2	β_4	β_{02}	√	√	√
		AGE_2 × SICDEGP1	β_5	β_{11}	√	√	√
		AGE_2 × SICDEGP2	β_6	β_{12}	√	√	√
		AGE_2SQ × SICDEGP1	β_7	β_{21}	√	√	
		AGE_2SQ × SICDEGP2	β_8	β_{22}	√	√	
Random effects	Child (i)	Intercept	u_{0i}	r_{0i}	√		
		AGE_2	u_{1i}	r_{1i}	√	√	√
		AGE_2SQ	u_{2i}	r_{2i}	√	√	√
Residuals	Time (t)		ε_{ti}	ε_{ti}	√	√	√
Covariance Parameters ($\boldsymbol{\theta}_D$) for $\boldsymbol{D}$ Matrix	Child level	Variance of intercepts	σ^2_{int}	τ[1,1]	√		
		Covariance of intercepts, AGE_2 effects	$\sigma_{int,age}$	τ[1,2]	√		
		Covariance of intercepts, AGE_2SQ effects	$\sigma_{int,age\text{-}squared}$	τ[1,3]	√		
		Variance of AGE_2 effects	σ^2_{age}	τ[2,2]	√	√	√
		Covariance of AGE_2 effects, AGE_2SQ effects	$\sigma_{age,age\text{-}squared}$	τ[2,3]	√	√	√
		Variance of AGE_2SQ effects	$\sigma^2_{age\text{-}squared}$	τ[3,3]	√	√	√
Covariance Parameters ($\boldsymbol{\theta}_R$) for $\boldsymbol{R}_i$ matrix	Time level	Residual variance	σ^2	σ^2	√	√	√

[a] The notation for the HLM software is described in more detail in Subsection 6.4.5.

Structures of Analyzing

- Data Summary
- General Model Specification
- Analysis Steps in R
- Hypothesis Tests
- Diagnostics for the final Model

Thank you