In the previous talks we have seen many definitions, algorithms,theorie about linear Mixed Models. Today in my talk we will see how to solve a problem systematically.

Overview of the talk 5

- Problem and Motivation
- Model-Building Strategies
- Two Examples
    1. Two-Level Models for Clustered Data : The Rat Pup Example
    2. Random Coefficient Models for Longitudinal Data:The Autism Example

Problem and Motivation

What is important in an application of LMM ?

- ✓ **Dependent Variable**: is actually the objective of our research, which is dependent on continuous or categorical covarates.
- ✓ **Covariates**: as fixed-effect parameters and random-effect parameters. fixed-effect parameters describe the relationships for an entire population, random effects are specific to clusters or subjects within a population, used in modeling the random variation at different levels of the data.
- ✓ **The relationships between a continuous dependent variable and various predictor variables**

Model-Building Strategies

The Goal of model selection is to choose the simplest model that best fit to the observed data. There are many possible choices. The top-down strategy will be used in our examples.

The Top-Down Strategy

1. **Start with a well-specified mean structure for the model:**Adding the fixed effects of covariates and interactions between the covariates, as many as possible .
2. **Select a structure for the random effects in the model:**Usually we choose some of the convarates in step 1. as random effects to include in the model.
3. **Select a covariance structure for the residuals in the model:**Once fixed effects and random effects have been added, we have residual error. An covariance structure for the residuals should be investigated
4. **Reduce the model**

The Rat Pup Study

30 female rats were separated into three groups, 10 rats each group, where they were treated with three different doses of an experimental compound (low, high, control dose). Then the weights and the sexes of their newborn pups were measured. 3 of the female rats in the high-dose group died, so there are no data for their litters.

This is an example of a two-level clustered data set, level 1 is about units of analysis (rat pups),Level 2 is about cluster (litter). The weights of the pups of the same litter are likely to be close to each other, because the mother of all the pups of a certain litter was exposed to the same environment and effects.

Data Summary

In addition, litter sizes varied widely, ranging from 2 to 18 pups. Because the number of litters per treatment and the number of pups per litter were unequal, the study has an unbalanced design.

Analysis Steps

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

Model 3.1 includes the fixed effects of treatment, sex, litter size, and the interaction between treatment and sex. The model also includes a random effect associated with the intercept for each litter and a residual associated with each birth weight observation. The residuals are assumed to be independent and identically distributed, with constant variance across the levels of treatment and sex.

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). Based on the result of this test,we decide to retain the random litter effects in all subsequent models.

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

Model 3.1: we take a homogeneous residual for all treatment groups.
Model 3.2A: we take a heterogeneous residual for each level of treatment(high, low, and control).
Model 3.2B, we take 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 (Model 3.2B, Model 3.3, or Model 3.3A).

We first test whether we wish to keep the treatment by sex interaction in Model 3.2B

Hypothesis Tests and Results

• Hypothesis 3.1: The random effects,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.

Random Coefficient Models for Longitudinal Data

we mean data sets in which the dependent variable is measured at several points in time for each unit of analysis.In another words: the observations are made on the same subject or unit of analysis over time. In our Autism Research example, Socialization score will be measured at some different ages

The Autism Example

214 children, who had been at autism clinics several times before the age of 3 years, were divided into three groups according to their language skills at the age of 2 years (autism, PDD,

nonspectrum). The children were then scored according to their VSAE (Vineland Socialization Age Equivalent ,including assessment of interpersonal relationships, play, coping skills) at the ages of 2, 3, 5, 9 and 13 years.

The data shows that mean VSAE scores generally increase with age. There may also be a quadratic trend in VSAE scores. Therefore a model to predict VSAE should include both linear and quadratic fixed effects of age and interactions between the both of them and the three groups.

Data Summary in R and Result of Data Summary

Figure 6.1:the VSAE scores of some children tend to increase as the children get older, for otherchildren remain relatively constant. At age 2 years ,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,

Figure 6.2:The mean VSAE scores generally increase with age. There may also be a quadratic trend in VSAE scores, especially in 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

We consider SICDEGP = 3 as the "reference category." When all covariates,including AGE_2, are equal to zero, the intercept can be interpreted as the VSAE score for children in the reference category of the SICD group (SICDEGP = 3).

Overview of the Autism Data Analysis

Hypothesis Tests

Analysis Steps in R

Step1: estimates of the parameters in Model 6.1 cannot be obtained using the summary()function. As a result, the model6.1.fit object is not created, and we proceed to consider Model 6.2 as an alternative.
Step2: Results from the fit of Model 6.2 are accessible using summary(model6.2.fit). we fit a nested model (Model 6.2A) by removing AGE_2SQ (specifically, I(age.2^2)) from the random portion of the syntax
Step3: Based on the p-value for the test of Hypothesis 6.2 (p = .39), we drop the fixed effects associated with interaction and obtain Model 6.3. An additional likelihood ratio test for the fixed effects associated with the age by SICD group interaction (i.e., Hypothesis 6.3) does not suggest that these tested fixed

Results of Hyothesis Tests

We test **Hypothesis 6.1** The significant (p < .001) means that the random effects associated with the quadratic (and therefore linear) effects of age should be retained in Model 6.2 and in all subsequent models.

We test **Hypothesis 6.2** using an ML-based likelihood ratio test. The test statistic is the value of the 2 ML log-likelihood for Model 6.3 (the nested model excluding the fixed effects associated with the interaction) minus the value for Model 6.2 (the reference model).To obtain a p-value for this statistic,

we refer it to a distribution with 2 degrees of freedom, corresponding to the 2 additional fixed-effect parameters in Model 6.2.

We test **Hypothesis 6.3** using an ML-based likelihood ratio test. The test statistic is the value of the 2 ML log-likelihood for Model 6.4 (the nested model excluding the fixed effects associated with the interaction) minus the value for Model 6.3 (the reference model). To obtain a p-value for this statistic, we refer it to a 2 distribution with 2 degrees of freedom, corresponding to the 2 additional fixed-effect parameters in Model 6.3. so we kept the

Diagnostics for the Final Model

Residual Diagnostics

Figure 6.6 The variance of the residuals appears to decrease for larger fitted values, and there are some possible outliers that may warrant further investigation. The preceding syntax may

Figure 6.7 suggests that the variance of the residuals is fairly constant across the values of AGE χ 2. We again note the presence of outliers.

Figure 6.8 suggests that the assumption of normality for the residuals seems acceptable. However, the presence of outliers in each level of SICDEGP (e.g., CHILDID = 46 in SICDEGP = 3) may warrant further investigation.

- **Diagnostics for the Random Effects**

Figure 6.9 We note that CHILDID = 124 is an outlier in terms of both random effects. The children indicated as outliers in these plots should be investigated in more detail to make sure that there is nothing unusual about their observations.The form of these plots is not suggestive of a very strong relationship between the random effects for age and age-squared

- **Observed and Predicted Values**

The distinguishing features of these plots are the outliers, which give the overall shape of the plots a rather unusual appearance. The EBLUPs for CHILDID = 124 are again unusual in Figure 6.10.

Figure 6.11 displays scatterplots of the observed VSAE scores vs. the conditional predicted VSAE scores for each level of SICDEGP. We see relatively good agreement between the observed and predicted values within each SICDEGP group, with the exception of some outliers.