Print

Summary: Lesson 6: Introduction to Linear Mixed Models

This summary contains topic summaries, syntax, and sample programs.

Topic Summaries

To go to the movie where you learned a task or concept, select a link.

Basics of Linear Mixed Models

To determine whether an effect is fixed or random, you look at the levels of that factor that are included in the study. An effect is fixed when the levels included in the study constitute the entire population in which you are interested. An effect is random when the levels included in the study represent only a sample from that population.

Fixed effects are factors whose levels are selected deliberately to evaluate the differences between those levels. All levels of interest are in the data set.

A random factor has a large number of levels and the researcher or data analyst selects a subset of the levels to be included in the study. This subset of levels represents a sample (although often an imperfect sample) from a population that has a probability distribution.

Let's review the general linear model. To fit a general linear model, you can use PROC GLM, PROC GLMSELECT, or PROC REG. The general linear model can also be called the fixed-effects-only linear model. Y is the vector of observed response data values. X is the known design matrix for the fixed effects, which is specified in the MODEL statement. β is the vector of unknown fixed-effect parameters. And ε is the vector of random errors. The fixed-effects-only linear model (that is, the general linear model) is actually a special case of the general linear mixed model.

In both the general linear model and the general linear mixed model, the dependent variable Y must have a normal distribution given the predictors. This means that the response must be a continuous variable. When the response has a nonnormal distribution given the predictors, the general linear mixed model is extended to the generalized linear mixed model. A response with a nonnormal distribution is typically, but not always, a categorical variable.

The general linear mixed model extends the general linear model in two ways. It includes random effects, and it allows a more flexible specification of the covariance matrix of the random errors. the linear mixed model allows for both correlated error terms and error terms with heterogeneous variances. The general linear mixed model contains an additional term, Zγ. Z is the known design matrix for the random effects. γ is the vector of unknown random-effect parameters.

Let's look at the three main assumptions that apply to general linear mixed models. The first assumption is that the random effects and random errors are normally distributed with a mean of zero and covariance matrices G and R, respectively.

The two sets of parameters in a linear mixed model actually specify the complete probability distribution of the data. The parameters of the mean model are referred to as fixed-effect parameters, that is, the betas in the term Xβ. The parameters of the variance-covariance model are referred to as covariance parameters. The G matrix describes the relationship among the gammas in the random-effects term Zγ. The R matrix describes the relationship among the errors in the epsilon term.

The second assumption for linear mixed models is that the random effects and random errors are independent of each other.

The third assumption is that the means (that is, the expected values) of the responses are linearly related to the predictor variables. In other words, the means are linear in terms of the fixed-effect parameters.

To fit a linear model with or without random effects, you can use PROC GLIMMIX. The PROC GLIMMIX syntax for fitting a linear mixed model includes the following statements: CLASS, COVTEST, LSMESTIMATE, MODEL, and two RANDOM statements – one with and one without the _RESIDUAL_ keyword.

The CLASS statement specifies the classification variables to be used in the analysis. The COVTEST statement provides a mechanism to obtain statistical inferences for the covariance parameters. The LSMESTIMATE statement requests custom hypothesis tests among the least squares means. The MODEL statement specifies a single response variable and the fixed effects.

The first RANDOM statement defines the Z matrix of the mixed model, the random effects in the γ vector, and the structure of G (the covariance matrix for the random effects). The second RANDOM statement (the one with the _RESIDUAL_ keyword) specifies the structure of R (the covariance matrix for the random errors).

```
PROC GLIMMIX <options>;
    CLASS variables;
    COVTEST <'label'> <test-specification> </ options>;
    LSMESTIMATE fixed-effect <'label'> values </ options>;
    MODEL response=<fixed-effects> </ options>;
    RANDOM random-effects> </ options>;
    RANDOM_RESIDUAL_ </ options>;
RUN;
```

If the variance of the random effect is contained in the G matrix, then it is called a G-side random effect. If the variance of the random effect is not an element of G, it is contained in the R matrix and is called an R-side random effect. R-side effects are also called residual effects. Models without G-side effects are known as marginal (or population-averaged) models.

For linear mixed models, PROC GLIMMIX uses the following estimation methods: To estimate variance and covariance parameters for normally distributed data, PROC GLIMMIX uses the restricted maximum likelihood method or the maximum likelihood method. To estimate fixed-effect parameters and standard errors for linear mixed models, PROC GLIMMIX uses the generalized least squares (or GLS) estimation method.

The GLS method requires knowledge of the G and R variance-covariance matrices and therefore is more appropriate for linear mixed models than the ordinary least squares method.

Fitting Linear Mixed Models

Data can be organized in two types of classification patterns, crossed or nested. When two factors are crossed, observations are collected for each combination of each level of the two factors. In nested classification, the factors are nested, that is, classified hierarchically.

For the school study, let's look at the the equation for the linear mixed model with nested classification and two factors: $y_{ijk} = \mu + \alpha_i + b(\alpha)_{ij} + \varepsilon_{ijk}$.

y_{ijk} represents the test score for the i^{th} material, the j^{th} teacher nested within the i^{th} material, and the k^{th} student within the j^{th} teacher for the i^{th} material, where i = 1 to 4, and j = 1 to 5, and k = 1 to 6.

μ represents the overall mean score.

α_i represents the fixed effects associated with the treatment variable, **Material**. Specifically, this term specifies the i^{th} material.

$b(\alpha)_{ij}$ represents the random effects associated with **Teacher** nested within **Material**, specifically, the j^{th} teacher nested within the ith material.

ε_{ijk} represents the random error associated with **Student**.

The random effects $b(\alpha)_{ij}$ are assumed to be independently and normally distributed with a mean of zero and variance σ_t^2. "Independently and normally distributed" is also referred to as "independently and identically distributed", i.i.d. The random errors are assumed to be independently and normally distributed with a mean of zero and variance σ^2. The effects $b(\alpha)_{ij}$ and ε_{ijk} are assumed to be independent random variables.

Let's consider hypothesis testing. In mixed model analysis, the hypotheses about the fixed effects are the same as those in the fixed-effects model, that is, whether there are significant treatment effects. The hypotheses

about the random effects are whether the variance components associated with the random effects equal zero; in other words, whether there are significant variations due to these random variables.

You can use the SGPANEL procedure to explore your data. PROC SGPANEL creates a panel of graph cells for the values of one or more classification variables. PROC SGPANEL also produces several types of layouts.

In the required PANELBY statement, you must specify one or more classification variables for the panel. The COLUMNS= option specifies the number of columns in the panel. The LAYOUT= option specifies one of four types of layouts for the panel: LATTICE, PANEL, COLUMNLATTICE, and ROWLATTICE.

The HISTOGRAM statement creates a histogram that displays the frequency distribution of a numeric variable. The HBOX statement creates a horizontal box plot that shows the distribution of your data. The REG statement creates a fitted regression line or curve. The SCATTER statement creates a scatter plot. The VBOX statement creates a vertical box plot that shows the distribution of your data.

PROC SGPANEL <*option(s)*>;
    PANELBY *variable(s) </ option(s)>*;
    HBOX *response-variable </ option(s)>*;
    HISTOGRAM *response-variable </ option(s)>*;
    REG X=*numeric-variable* **Y=***numeric-variable </ option(s)>*;
    SCATTER X=*variable* **Y=***variable </ option(s)>*;
    VBOX *response-variable </ option(s)>*;
RUN;

Sample Programs

Fitting a Linear Mixed Model with Nested Classification

```
title1 "Distribution of Scores by Materials";
proc sgpanel data=mydata.scores;
   panelby material / columns=4;
   vbox score;
run;

proc glimmix data=mydata.scores;
   class material teacher;
   model score=material;
   random teacher(material);
   covtest 'Test Need for Random Effect' glm;
run;

title 'Random Effect is Incorrectly Specified as Fixed Effect';
proc glimmix data=mydata.scores;
   class material teacher;
   model score=material teacher(material);
   output out=checkvar variance=ResidualVariance;
run;

proc print data=checkvar (obs=1);
   var ResidualVariance;
run;

title;
```

Close