BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 10

Synthesis of Ideas for Analyzing Longitudinal Data

Synthesis of Ideas for Analyzing Longitudinal Data

Primary goal of a longitudinal study is to characterize the change in response over time and the factors that influence change.

Longitudinal data require somewhat more sophisticated statistical techniques because: (i) repeated measures on the same individual are usually positively correlated, and (ii) variability is often heterogeneous across measurement occasions.

Correlation and heterogeneous variability must be accounted for in order to obtain valid inferences about change in response over time.

General Linear Model for Longitudinal Data

So far, we have considered linear regression models that

- permit individuals to be measured on different number of occasions and at different times
- can handle mixed discrete and continuous covariates
- allow a range of different covariance structures

We consider linear regression models for changes in the mean response over time:

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + e_{ij}, \;\; j = 1, ..., n_i;$$

where $\beta_1, ..., \beta_p$ are unknown regression coefficients.

The e_{ij} are random errors, with mean zero, and represent deviations of the Y_{ij}'s from their means,

$$E(Y_{ij}|X_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}.$$

This model can also be represented in vector/matrix notation as:

$$E(Y_i|X_i) = X_i\beta.$$

Assumptions

(1) The individuals represent a random sample from the population of interest.

(2) Observations from different individuals are independent, while repeated measurements of the same individual are not assumed to be independent.

(3) The elements of the vector of repeated measures $Y_{i1}, \ldots, Y_{in_i}$, have a Multivariate Normal (MVN) distribution, with means

$$\mu_{ij} = E(Y_{ij}|X_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$$

and covariance matrix Σ_i[1].

(4) If there are missing data they are assumed to be "missing at random" (MAR) or "missing completely at random" (MCAR). (Lecture 21)

[1] Covariance matrix is indexed by i to permit individuals to have different numbers of repeated measures, n_i

Modelling Longitudinal Data

We have stressed that, in fitting linear models to longitudinal data, we have two modeling tasks:

1. We must choose a covariance model that provides a good fit to the observed variances and covariances.

2. We must fit a linear regression model that provides a good fit to the mean of the outcome variable.

Because the models for the mean and covariance are interdependent, we need to have a coherent strategy for model fitting.

Choosing a Covariance Structure

The choices of models for the mean and covariance are interdependent.

Since the residuals depend on the specification of the linear model for the mean, we choose a covariance structure for a particular linear model.

Substantial changes in the linear model could lead to a different choice of model for the covariance.

A balance needs to be struck:

With too little structure (e.g., unstructured), there may be too many parameters to be estimated with the limited amount of data available.
This would leave too little information available for estimating β

$\Rightarrow$ weaker inferences concerning β.

With too much structure (e.g., compound symmetry), there is more information available for estimating β.
However, there is a potential risk of model misspecification

$\Rightarrow$ apparently stronger, but potentially biased, inferences concerning β.

General Strategy for Model Fitting

(1) To analyze longitudinal data we first need to choose a "working" covariance structure.

We must recognize that choices of model for the mean and covariance are interdependent.

Need to fit a "maximal model" for the mean response when choosing/comparing models for the covariance.

Can use REML log likelihood or AIC as criteria to guide the choice of model for the covariance.

When n is relatively small, and design is balanced, can simply use unstructured covariance matrix unless simpler model is clearly satisfactory.

When n is relatively large and/or there are mistimed measurements, alternative models for the covariance will need to be considered.

(2) Given choice of "working" covariance, select model for mean response. Need to decide how to model the pattern of change in the mean response:

(a) covariate by time interaction(s), where time is regarded as a categorical variable (analysis of response profiles)

(b) covariate by time interaction(s), where means are modeled as an explicit function of continuous time (parametric and semi-parametric curves)

(c) covariate effects in an analysis that includes the baseline measure as a covariate (e.g., randomized study)

(d) covariate by post-baseline time (posttime) interaction(s) in a "constant effect" model

Use the ML log likelihood to compare nested models for the mean differing by several degrees of freedom.

(3) Make an initial determination of the final form of the regression model.

(4) If necessary, re-fit the final regression model using REML to obtain standard errors.

Empirical Variance Estimation

We have focused on regression models for longitudinal data where the primary interest is in making inference about the regression parameters β.

For statistical inference about β we need

(i) an estimate, $\widehat{\beta}$

(ii) estimated standard error, $\text{SE}(\widehat{\beta})$

So far, we have made inferences about β using standard errors obtained under an assumed model for the covariance structure.

This approach is potentially problematic if the assumed covariance has been mis-specified.

How might the covariance be mis-specified?

For example, compound symmetry might be assumed but the correlations in fact decline over time.

Alternatively, an unstructured covariance might be assumed but the covariances also depend upon the treatment group.

If the assumed covariance has been mis-specified, we can correct the standard errors by using "empirical" or so-called "robust" variances.

Recall, the REML estimator of β is given by

$$\widehat{\beta} = \left[\sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} X_i\right)\right]^{-1} \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} Y_i\right)$$

where $\widehat{\Sigma}$ is the REML estimate of Σ.

It has covariance matrix,

$$\text{Cov}(\widehat{\beta}) = \left[\sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} X_i\right)\right]^{-1} \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} \text{Cov}\left(Y_i\right) \widehat{\Sigma}^{-1} X_i\right) \left[\sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} X_i\right)\right]^{-1}$$

If $\text{Cov}(Y_i)$ is replaced by $\widehat{\Sigma}$, the REML estimate of Σ, $\text{Cov}(\widehat{\beta})$ can be estimated by

$$\left[\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}$$

However, if the covariance has been mis-specified then an alternative estimator for $\text{Cov}\,(Y_i)$ is needed.

The empirical or so-called robust variance of $\widehat{\beta}$ is obtained by using

$$\widehat{V}_i = \left(Y_i - X_i\widehat{\beta}\right)\left(Y_i - X_i\widehat{\beta}\right)'$$

as an estimate of $\text{Cov}\,(Y_i)$.

Thus, the empirical variance of $\widehat{\beta}$ is estimated by

$$\left[\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}\widehat{V}_i\widehat{\Sigma}^{-1}X_i\right)\left[\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}$$

This empirical variance estimator is also known as the "sandwich estimator".

The remarkable thing about the empirical estimator of $\text{Cov}(\widehat{\beta})$ is that it provides a consistent estimator of the variance even when the model for the covariance matrix has been misspecified.

That is, in large samples the empirical variance estimator yields correct standard errors.

In general, its use should be confined to cases where N (number of individuals) is relatively large and n (number of measurements) is relatively small.

The empirical variance estimator may not be appropriate when there is severe imbalance in the data.

In summary, (with large samples) the following procedure will produce valid estimates of the regression coefficients and their standard errors:

(1) Choose a "working" covariance matrix of some convenient form.

(2) Estimate the regression coefficients under the assumed working covariance matrix.

(3) Estimate the standard errors using the empirical variance estimator.

Why not be a clever ostrich?

Why not simply ignore potential correlation among repeated measures (i.e., put head in sand) and assume an independence "working" covariance? Then, obtain correct standard errors using empirical variance estimator. Why should we bother to explicitly model the covariance?

Reasons:

(1) Efficiency: The optimal (most precise) estimator of β uses the true $\text{Cov}(Y_i)$. Given sufficient data, we can attempt to estimate $\text{Cov}(Y_i)$.

(2) When N (number of individuals) is not large relative to n (number of measurements) the empirical variance estimator is not recommended.

(3) Missing values: The empirical variance estimator uses the replications across individuals to estimate the covariance structure. This becomes problematic when there are missing data or when the times of measurement are not common.

In general, it is advantageous to model the covariance.

Table 1: Illustrative commands for an exponential model,
with empirical standard errors, using PROC MIXED in SAS.

```
PROC MIXED EMPIRICAL;
  CLASS id group time;
  MODEL y=group time group*time /S CHISQ;
  REPEATED time / TYPE=SP(EXP)(ctime) SUBJECT=id R RCORR;
```


BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 11

Mixed Effects Models for Longitudinal Data

Two-Stage (Two-Level) Random Effects Formulation

Two-Stage Random Effects Formulation: Example

Feldman (1988) describes a small animal study designed to compare clearance of iron particles from the lung and liver.

Iron oxide particles were administered to four rats by intravenous injection and to four other rats by tracheal installation.

The injected particles were taken up by liver endothelial cells and the installed particles by lung macrophages.

Each rat was followed for 30 days, during which time the quantity of iron oxide remaining in the lung was measured by magnetometry.

The iron oxide content declined linearly on the logarithmic scale.

The goal of the study was to compare the rate of particle clearance by liver endothelial cells and by lung macrophages.

Follow-up measurements were expressed as a percentage of the baseline value (so the baseline value is not used in the analysis).

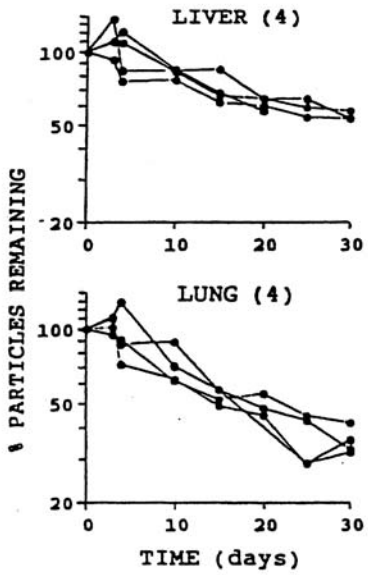

Figure 1: Timeplot of Feldman's clearance data (note log scale).

Two-Stage (Two-Level) Formulation

Linear mixed effects models can be motivated in terms of the following two-stage formulation of the model.

Basic idea: In the two-stage formulation of the model, we assume

1. A straight line (or more generally a "growth" curve) fits the observed responses for each subject (**first stage or level**)

2. A regression model relating the mean of the individual intercepts and slopes to subject-specific covariates (**second stage or level**)

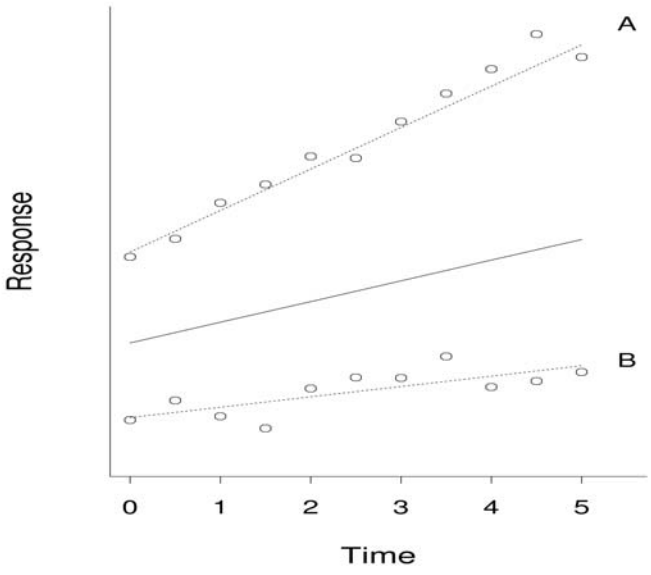

Figure 2: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

Stage 1

In the first stage, each subject is assumed to have their own unique individual-specific mean response trajectory. So for subject i:

$$Y_{ij} = Z'_{ij}\beta_i + \epsilon_{ij}, \quad (j = 1, ..., n_i)$$

where β_i is a vector of subject-specific regression parameters; the errors, ϵ_{ij}, are (usually) assumed to be independent within a subject.

For example, a simple model with subject-specific intercepts and slopes over time is given by

$$Y_{ij} = \beta_{1i} + \beta_{2i}t_{ij} + \epsilon_{ij}.$$

Thus, in stage 1 we posit a regression model with separate or distinct coefficients for each subject.

This is equivalent to considering separate linear regression models for the data for each subject.

Note: Covariates in Z_{ij} are restricted to covariates that vary within the subject over time (i.e. time-varying covariates), except for the column of 1's for the intercept.

Time-invariant or between-subject covariates (e.g., gender, treatment group, exposure group) cannot be included in Z_{ij}; instead, they are introduced in the second stage of the model formulation.

Stage 2

In the second stage, we assume that the subject-specific effects, the β_i's, are random (i.e. there is some distribution for the β_i's in the population, e.g. a normal distribution).

The mean and covariance of the β_i's are the population parameters that are modelled in the second stage.

Specifically, variation in the β_i's in the population is described in terms of between-subject covariates, say A_i (e.g., gender, treatment group):

$$\beta_i = A_i\beta + b_i, \text{ where } b_i \sim N(0, G).$$

For example, consider two-group (e.g. treatment vs. control) setting and the simple model with subject-specific intercepts and slopes.

Allowing both the mean intercept and slope to depend on group

$$\begin{aligned} E(\beta_{1i}) &= \beta_1 + \beta_2 \, \texttt{Group}_\texttt{i} \\ E(\beta_{2i}) &= \beta_3 + \beta_4 \, \texttt{Group}_\texttt{i} \end{aligned}$$

where $\texttt{Group}_\texttt{i} = \texttt{1}$ if the i^{th} individual was assigned to the treatment, and $\texttt{Group}_\texttt{i} = \texttt{0}$ otherwise.

In this model, β_1 is the mean intercept in the control group, while $\beta_1 + \beta_2$ is the mean intercept in the treatment group.

Similarly, β_3 is the mean slope in the control group, while $\beta_3 + \beta_4$ is the mean slope in the treatment group.

In this model, the design matrix A_i of between-subject covariates has the following form:

$$A_i = \begin{pmatrix} 1 & \texttt{Group}_\texttt{i} & 0 & 0 \\ 0 & 0 & 1 & \texttt{Group}_\texttt{i} \end{pmatrix}.$$

Thus, for the control group, the model for the mean is

$$E \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix};$$

and similarly, for the treatment group, the model for the mean is

$$E \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_2 \\ \beta_3 + \beta_4 \end{pmatrix}.$$

It is also assumed that there is residual variation in the β_i's, that cannot be explained by the effect of group.

As $\beta_i = A_i\beta + b_i$, this is the variability in the b_i's given by

$$\text{Cov}(\beta_i|A_i) = \text{Cov}(b_i) = G = \begin{pmatrix} g_{11} & g_{12} \\ \\ g_{21} & g_{22} \end{pmatrix},$$

where $g_{11} = \text{Var}(b_{1i})$, $g_{22} = \text{Var}(b_{2i})$, and $g_{12} = g_{21} = \text{Cov}(b_{1i}, b_{2i})$.

Thus, g_{11} is the variance of β_{1i}, after adjusting for the effect of treatment group, and so on.

We can combine the two components of the two-stage model:

$$\begin{aligned} Y_{ij} &= Z'_{ij}\beta_i + \epsilon_{ij} \\ &= Z'_{ij}(A_i\beta + b_i) + \epsilon_{ij} \\ &= (Z'_{ij}A_i)\beta + Z'_{ij}b_i + \epsilon_{ij} \\ &= X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij}, \end{aligned}$$

where $X'_{ij} = Z'_{ij}A_i$.

$\Longrightarrow$ Linear Mixed Effects Model (albeit with constraint, $X'_{ij} = Z'_{ij}A_i$).

Mixed Effects Model Representation: Linear Trend

We can develop a mixed effects model in two stages corresponding to the two-stage model:

Stage 1:

$$Y_{ij} = \beta_{1i} + \beta_{2i} t_{ij} + \epsilon_{ij}$$

where β_{1i} and β_{2i} are the intercept and slope for the i^{th} subject,

and the errors, ϵ_{ij}, are assumed to be independent and normally distributed around the individual's regression line, i.e. $\epsilon_{ij} \sim N\left(0, \sigma^2\right)$.

Stage 2:

Assume that the intercept and slope, β_{1i} and β_{2i}, are random and have a joint multivariate normal distribution, with mean dependent on covariates (e.g., the organ studied):

$$\beta_{1i} = \beta_1 + \beta_2 \text{ Organ } + b_{1i}$$

$$\beta_{2i} = \beta_3 + \beta_4 \text{ Organ } + b_{2i}$$

Also, let $(b_{1i}) = g_{11}$, $(b_{1i}, b_{2i}) = g_{12}$, $(b_{2i}) = g_{22}$.

If we substitute the expressions for β_{1i} and β_{2i} into the equation in stage 1, we obtain

$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2 \text{ Organ } + \beta_3 t_{ij} + \beta_4 \text{ Organ } \times t_{ij} \\ &\quad + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij} \end{aligned}$$

The first four terms give the regression model for the mean response implied by the two-stage model (the fixed effect part of the model).

The last three terms are the random "error terms" (between- and within-subject).

Two-Stage Analysis: "NIH Method"

One classic approach with a long history for the analysis of longitudinal data is known as two-stage or two-step analysis; sometimes called the NIH Method because it was popularized by statisticians working at NIH.

In the two-stage analysis method, we fit a straight line (or curve) to the $\log_{10}$ response data for each subject (stage 1), and then regress the estimates of the individual intercepts and slopes on subject-specific covariates (stage 2).

One of the attractions of this method is that it was very easy to perform using existing statistical software for standard linear regression (i.e., before software like PROC MIXED).

The two-stage analysis is easy to understand and nearly efficient when the dataset is balanced and complete.

It is somewhat less attractive when the number and timing of observations varies among subjects, because it does not take proper account of the different precisions among subjects in estimating their intercepts and slopes.

In contrast, we can consider the mixed effects model corresponding to the two-stage model, and obtain efficient (more precise) estimates of the regression coefficients.

Linear Mixed Effects Model Induced by the Two-Stage Formulation

$$Y_{ij} \quad = \quad X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij}, \text{ where } X'_{ij} = Z'_{ij}A_i.$$

For the vector of responses, Y_i, this becomes:

$$Y_i \quad = \quad X_i\beta + Z_ib_i + \epsilon_i, \text{ where } X_i = Z_iA_i.$$

Note that $E(Y_i|X_i) = X_i\beta$ is the marginal mean, and β is the vector of fixed effects.

$$Y_i \quad = \quad X_i\beta + Z_i b_i + \epsilon_i, \text{ where } X_i = Z_i A_i.$$

b_i and ϵ_i are (independent) random effects (between- and within-subject "errors")

There is an induced variance-covariance structure:

$$\text{Cov}(Y_i) = Z_i G Z_i{}' + \sigma^2 I_{n_i}.$$

In the simple model with random intercepts and slopes, this gives:

$$\begin{array}{lcl} \text{Var}(Y_{ij}) & = & g_{11} + 2t_{ij}g_{12} + t_{ij}^2 g_{22} + \sigma^2, \\ \text{Cov}(Y_{ij}, Y_{ik}) & = & g_{11} + 2(t_{ij} + t_{ik})g_{12} + t_{ij}t_{ik}g_{22} \end{array}$$

PROC MIXED in SAS

This model can be fit using the ***RANDOM*** statement in PROC MIXED.

```
FILENAME rats 'g:\shared\bio226\rat.txt';

DATA clear;
     INFILE rats;
     INPUT organ $ id days cfp logcfp;
     IF (days=0) THEN DELETE;
RUN;

PROC MIXED DATA=clear;
     CLASS id organ;
     MODEL logcfp=days organ days*organ / S CHISQ;
     RANDOM INTERCEPT days /
          TYPE=UN SUBJECT=ID G;
     TITLE 'Random Slopes and Intercepts';
RUN;
```

Random Slopes and Intercepts

Estimated G Matrix

Parameter	ID	Row	col1	col2
Intercept	1	1	0.002851	-0.00015
days	1	2	-0.00015	9.65E-6

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
UN(1,1)	ID	0.002851
UN(2,1)	ID	-0.00015
UN(2,2)	ID	9.65E-6
Residual		0.003155

Random Slopes and Intercepts
Solution for Fixed Effects

Effect	organ	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		2.0375	0.03337	6	61.05	<.0001
days		-0.01785	0.001913	6	-9.33	<.0001
organ	liver	0.003814	0.04741	37	0.08	0.9363
organ	lung	0	.	.	.	.
days*organ	liver	0.006232	0.002760	37	2.26	.0299
days*organ	lung	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	Pr > ChiSq
days	1	6	114.05	<.0001
organ	1	37	0.01	0.9359
days*organ	1	37	5.10	0.0239

Results suggest that mean clearance of foreign particles is faster from the lung.

Estimated slope in the lung group is -0.0178, representing a half time for clearance of 16.9 days (or $\frac{\log_{10}(0.5)}{-0.0178}$).

Estimated slope in the liver group is -0.0116 (-0.0178 + 0.0062), representing a half time for clearance of 26.0 days.

The mean slopes in the two groups are different ($p < 0.05$).

Side-by-Side Comparison of Results

	Two-Stage		**Combined**	
Intercept	2.0375	(.0353)	2.0375	(.0334)
Day	-0.0178	(.0022)	-0.0179	(.0019)
Organ	0.0104	(.0450)	0.0038	(.0474)
Organ*Day	0.0054	(.0031)	0.0062	(.0027)

Comments

The two-stage analysis method is less attractive because it is less efficient, particularly when the number and timing of observations varies among subjects (because it does not take proper account of the relative precisions of the estimated β_i's).

Also, the two-stage formulation of the growth curve model imposes certain restrictions and structure on the covariates: those at the first stage (except for the intercept) must be *time-varying*, while those at the second stage must be *time-invariant*.

In contrast, there is a more general formulation of the mixed effects model in which the only restriction is that the components of Z_{ij} are a subset of the components of X_{ij}.