BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 5

Statistical Basis of Longitudinal Analysis (Part 2)

Statistical Basis of Longitudinal Analysis (Part 2)

Overview:

Previously, we introduced some additional vector and matrix notation.

We also presented a general linear regression model for longitudinal data:

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

Next, we consider distributional assumptions and discuss inference based on maximum likelihood (ML).

General Linear Model for Longitudinal Data

We assume a general linear regression model,

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

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

The e_{ij} are random errors, with mean zero, and are expected to be correlated within individuals.

That is, $\text{Cov}(e_{ij}, e_{ij'}) \neq 0 \quad (j \neq j')$.

To simplify notation, in the following we assume that $n_i = n$.

Assumptions

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

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

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

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

The covariance matrix of the vector of observations, $Y_{i1}, \ldots, Y_{in}$, is denoted Σ and its elements are $\sigma_{jj'}$ (typically, we denote variances, σ_{jj}, by σ_j^2).

Probability Models

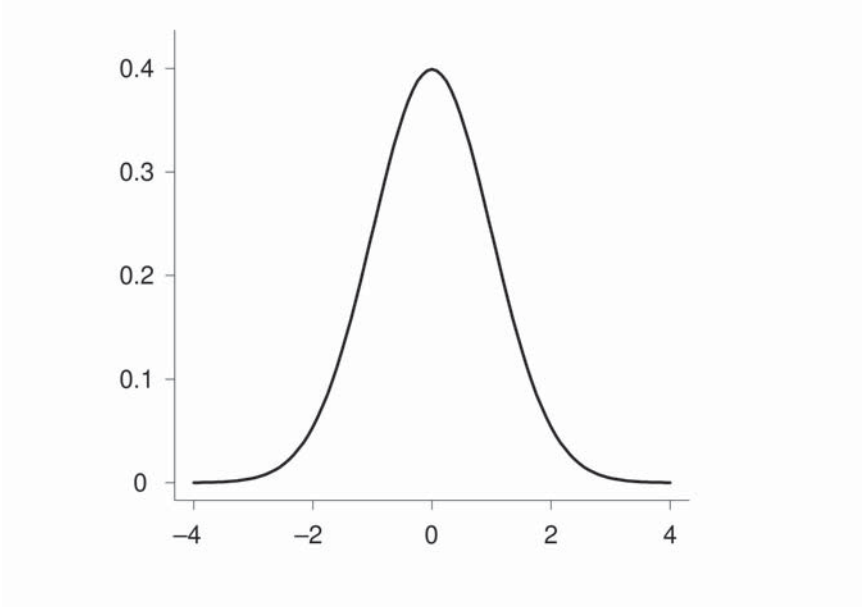
The foundation of most statistical procedures is a probability model, i.e., probability distributions are used as models for the data.

A probability distribution describes the probability or relative frequency of occurrence of particular values of the response (or dependent) variable.

Recall: The normal probability density for a single response variable, say Y_i, in the standard linear regression model is:

$$f(y_i|X_{i1},...,X_{ip}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i - [\beta_1 X_{i1} + \cdots + \beta_p X_{ip}])^2}{2\sigma^2}\right\}$$

Equivalently, we assume that $e_i \sim N(0, \sigma^2)$



With repeated measures we have a vector of response variables and must consider joint probability models for the entire vector of responses.

A joint probability distribution describes the probability or relative frequency with which the vector of responses takes on a particular set of values.

The Multivariate Normal Distribution is an extension of the Normal distribution for a single response to a vector of responses.

Multivariate Normal Distribution

The multivariate normal probability density function for $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$ has the following representation:

$$f(Y_i|X_i) = f(Y_{i1}, Y_{i2}, \ldots, Y_{in}|X_i) =$$

$$(2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left[-(Y_i - X_i\beta)' \Sigma^{-1} (Y_i - X_i\beta)/2\right]$$

where $|\Sigma|$ is the *determinant* of Σ (also known as the *generalized variance*).

Note that $f(Y_i|X_i)$ describes the probability or relative frequency of occurrence of a particular set of values of $(Y_{i1}, Y_{i2}, \ldots, Y_{in})$.

Notable Features:

- $f(Y_i|X_i)$ is completely determined by the vector of means, $\mu_i = X_i\beta$, and by Σ
- $f(Y_i|X_i)$ depends to a very large extent on

$$(Y_i - X_i\beta)' \Sigma^{-1} (Y_i - X_i\beta)$$

- Although somewhat more complicated than in the univariate case, the latter has interpretation in terms of a measure of distance

In the *bivariate* case, it can be shown that

$$(Y_i - \mu_i)' \Sigma^{-1} (Y_i - \mu_i) =$$

$$(1-\rho^2)^{-1} \left\{ \frac{(Y_{i1}-\mu_1)^2}{\sigma_{11}} + \frac{(Y_{i2}-\mu_2)^2}{\sigma_{22}} - 2\rho \frac{(Y_{i1}-\mu_1)(Y_{i2}-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}} \right\}$$

where $\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$.

Note that this measure of *distance*

(i) down-weights deviations from the mean when the variance is large; this make intuitive sense because when the variance is large the "information" is somewhat poorer; and

(ii) modifies the distance depending on the magnitude of the correlation; when there is strong correlation, knowing that Y_{i1} is "close" to μ_1 also tells us something about how close Y_{i2} is to μ_2.

Maximum Likelihood and Generalized Least Squares

Next we consider a framework for estimation of the unknown parameters, β and Σ.

When full distributional assumptions have been made for vector of responses a standard approach is to use the method of *maximum likelihood* (ML).

Recall main idea behind ML: use as estimates of β and Σ the values that are most probable (or "likely") for the data that we have observed.

That is, choose values of β and Σ that maximize the probability of the response variables evaluated at their observed values.

Regression with Independent Observations

For standard linear regression, with independent observations, the joint density is the product of the individual univariate normal densities.

We maximize

$$\prod_{i=1}^{N} f(y_i|X_{i1},...,X_{ip}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2}\right\}$$
$$= \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\sum_{i=1}^{N} \frac{(Y_i - X_i'\beta)^2}{2\sigma^2}\right\},$$

with respect to the regression parameters, β,
or minimize

$$\sum_{i=1}^{N} (Y_i - X_i'\beta)^2 / 2\sigma^2$$

Generalized Least Squares

To find ML estimate of β in the repeated measures setting we first assume Σ is known (later, we will relax this unrealistic assumption).

Given that $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$ are assumed to have a multivariate normal distribution, we must maximize the following log-likelihood

$$\begin{aligned}
&\ln \{ (2\pi)^{-Nn/2} |\Sigma|^{-N/2} \\
&\exp \left[-\sum_{i=1}^{N} (Y_i - X_i\beta)' \Sigma^{-1} (Y_i - X_i\beta) /2 \right] \} \\
&= -\tfrac{Nn}{2} \ln (2\pi) - \tfrac{N}{2} \ln |\Sigma| \\
&- \left[\sum_{i=1}^{N} (Y_i - X_i\beta)' \Sigma^{-1} (Y_i - X_i\beta) /2 \right]
\end{aligned}$$

or minimize

$$\sum_{i=1}^{N} (Y_i - X_i\beta)' \Sigma^{-1} (Y_i - X_i\beta)$$

The estimate of β that minimizes this expression is known as the generalized least squares (GLS) estimate and can be written as

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

This is the estimate that PROC MIXED in SAS provides.

Properties of GLS

(1) For any choice of Σ, GLS estimate of β is unbiased; that is, $E(\widehat{\beta}) = \beta$.

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

(3) Sampling Distribution of $\widehat{\beta}$:

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

The most efficient generalized least squares estimate is the one that uses the true value of Σ.

Since we usually do not know Σ, we typically estimate it from the data.

In general, no simple expression for ML estimate of Σ.

It has to be found using numerical algorithms that maximize the likelihood.

Once ML estimate of Σ, say $\widehat{\Sigma}$, has been obtained, we substitute it in the GLS estimator to obtain ML estimate of β,

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

In large samples, resulting estimator of β has all the same properties as when Σ is known.

Residual Maximum Likelihood (REML) Estimation

Recall: ML estimate of β and Σ is obtained by maximizing the following log-likelihood

$$-\tfrac{Nn}{2}\ln(2\pi) - \tfrac{N}{2}\ln|\Sigma|$$

$$-\left[\sum_{i=1}^{N}(Y_i - X_i\beta)'\,\Sigma^{-1}\,(Y_i - X_i\beta)/2\right]$$

Although the MLEs have the usual large sample (or asymptotic) properties, the MLE of Σ has well-known bias in small samples (e.g., the diagonal elements of Σ are underestimated).

To see problem, consider linear regression with independent errors.

If the N observations are independent we maximize

$$\prod_{i=1}^{N} f(y_i|X_{i1},...,X_{ip}) = \left(2\pi\sigma^2\right)^{-N/2}\exp\left\{-\sum_{i=1}^{N}\frac{(Y_i - X_i'\beta)^2}{2\sigma^2}\right\}.$$

This gives the usual least squares estimator of β, but ML estimator of σ^2 is

$$\widehat{\sigma}^2 = \sum_{i=1}^{N}\left(Y_i - X_i'\widehat{\beta}\right)^2/N$$

Note: The denominator is N. Furthermore, it can be shown that

$$E(\widehat{\sigma}^2) = \left(\frac{N-p}{N}\right)\sigma^2.$$

As a result, the ML estimate of σ^2 will be biased in small samples and will underestimate σ^2.

In effect, the bias arises because the ML estimate has not taken into account that β, also, is estimated. That is, in the estimator of σ^2 we have replaced β by $\widehat{\beta}$.

It should not be too surprising that similar problems arise in the estimation of Σ.

Recall: An unbiased estimator is given by using $N - p$ as the denominator instead of N.

The theory of residual or restricted maximum likelihood estimation was developed to address this problem.

The main idea behind REML is to eliminate the parameters β from the likelihood so that it is defined only in terms of Σ.

One possible way to obtain the restricted likelihood is to consider transformations of the data to a set of linear combinations of observations that have a distribution that does not depend on β.

For example, the residuals after estimating β by ordinary least squares can be used.

The likelihood for these residuals will depend only on Σ, and not on β.

Thus, rather than maximizing

$$-\frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - X_i\widehat{\beta}\right)'\Sigma^{-1}\left(Y_i - X_i\widehat{\beta}\right)$$

REML maximizes the following slightly modified log-likelihood

$$\begin{aligned} -\frac{N}{2}\ln|\Sigma| \quad &- \quad \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - X_i\widehat{\beta}\right)'\Sigma^{-1}\left(Y_i - X_i\widehat{\beta}\right) \\ &- \quad \frac{1}{2}\ln\left|\sum_{i=1}^{N}X_i'\Sigma^{-1}X_i\right| \end{aligned}$$

When the residual likelihood is maximized, we obtain less biased estimate of Σ.

That is, the extra determinant term effectively makes a correction or adjustments that is analogous to the correction to the denominator in $\hat{\sigma}^2$.

When REML estimation is used, we obtain the GLS estimates of β,

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

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

In PROC MIXED, REML is the default maximization criterion.

ML estimates are obtained by specifying:

PROC MIXED METHOD = ML;

Statistical Inference: Likelihood Ratio Test

Suppose that we are interested in comparing two *nested* models, a "full" model and a "reduced" model.

Aside: Nested Models

When one model (the "reduced" model) is a special case of the other (the "full" model), the reduced model is said to be *nested* within the full model.

We can compare two nested models by comparing their maximized log-likelihoods, say $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$; the former is at least as large as the latter.

The larger the difference between $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$ the stronger the evidence that the reduced model is inadequate.

A formal test is obtained by taking

$$2(\widehat{l}_{\text{full}} - \widehat{l}_{\text{red}})$$

and comparing the statistic to a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models.

Formally, this test is called the *likelihood ratio test* (LRT).

We can use LRTs for hypotheses about models for the mean and the covariance[1].

[1] Later in the course, we will discuss some potential problems with the use of the likelihood ratio test for comparing nested models for the covariance.

Note: The residual maximum likelihood (REML) can be used to compare different models for the covariance structure.

However, it should <u>not</u> be used to compare different mean functions (i.e. involving the β) in the regression models since the penalty term in REML depends upon the regression model specification.

Recall: REML maximizes the following slightly modified log-likelihood

$$\begin{aligned} -\frac{N}{2}\ln|\Sigma| &- \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - X_i\widehat{\beta}\right)'\Sigma^{-1}\left(Y_i - X_i\widehat{\beta}\right) \\ &- \frac{1}{2}\ln\left|\sum_{i=1}^{N}X_i'\Sigma^{-1}X_i\right| \end{aligned}$$

Instead, the standard ML log-likelihood should be used for comparing different regression models for the mean.

Reminder: In PROC MIXED, REML is the default maximization criterion.

ML estimates are obtained by specifying:

PROC MIXED METHOD = ML;