BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 13

Aspects of Design of Longitudinal Studies

Aspects of Design of Longitudinal Studies

In this lecture, we consider two issues concerning the design of longitudinal studies:

(1) Sample Size and Power

(2) Longitudinal and Cross-Sectional Information

The first issue has important implications for planning a longitudinal study, the second has implications for analysis.

Sample Size and Power

Investigators typically need to know the answer to the following question: "How *large* should my study be?"

Answer is straightforward with only a single, univariate response: the *size* of a study = sample size.

For a longitudinal study the question of *size* is more complex,
e.g., number of subjects, duration of study, frequency and spacing of repeated measurements on the subjects.

Before discussing sample size/power in context of longitudinal studies, we review sample size/power formulas for a univariate response.

We then describe a simple, albeit approximate, method that allows direct application of these formulas in longitudinal setting.

Sample Size for a Univariate Response

Interested in comparing two treatments (or exposures), denoted A and B.

Plan to randomize an equal number of subjects, say N, to two groups.

Two groups are to be compared in terms of the mean response.

Let $\mu^{(A)}$ and $\mu^{(B)}$ denote the mean response in the two populations.

Define effect of interest as $\delta = \mu^{(A)} - \mu^{(B)}$.

The null hypothesis of no group difference is H_0: $\delta = 0$.

Type I and Type II Errors

Recall: Two types of errors can arise when testing H_0: $\delta = 0$.

Type I error: If we reject the null hypothesis when in fact it is true.

Thus, for our example where H_0: $\delta = 0$,

$$\alpha = \Pr(\text{Reject } H_0 \mid H_0 \text{ is true}).$$

The probability of type I error, also known as the significance level, is usually denoted by α.

Often, α is chosen to be no greater than 0.05.

Type II error: If we fail to reject the null hypothesis when in fact it is false.

We denote probability of a type II error by γ, with

$$\gamma = \Pr(\text{Fail to reject } H_0 \mid H_0 \text{ is false}).$$

Note: γ necessarily depends upon the particular choice of value for $\delta \neq 0$ under the alternative hypothesis.

Finally, power of test is defined as $1 - \gamma$, that is,

$$\text{power} = 1 - \gamma = \Pr(\text{Reject } H_0 \mid H_0 \text{ is false}).$$

Two-Group Sample Size Formula

For two group comparison, a formula for approximate sample size in each group, N, is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\,\sigma^2}{\delta^2}, \text{ where}$$

σ^2 is variance of response (assumed to be common in two groups), and $Z_{(1-\alpha/2)}$ and $Z_{(1-\gamma)}$ denote the $(1-\alpha/2) \times 100\%$ and $(1-\gamma) \times 100\%$ percentiles of a standard normal distribution (e.g., the 97.5th percentile of a standard normal distribution is 1.96).

Note: N denotes sample size in each group; total sample size is $2N$.

Closer examination of formula reveals that the determination of sample size requires

(i) significance level, α;

(ii) power, $1-\gamma$;

(iii) effect size, δ; and

(iv) common variance, σ^2.

Often, α is fixed at 0.05 level (with $Z_{(1-\alpha/2)} = 1.96$ for a 2-tailed test) though this choice is somewhat arbitrary.

Similarly, lower bound on acceptable power is usually set at 80% or 90% (with $Z_{(1-\gamma)} = 0.842$ for power $= 0.8$, or $Z_{(1-\gamma)} = 1.282$ for power $= 0.9$).

Investigators must provide information on: minimum effect size of scientific interest, δ, and an estimate of σ^2.

Sample Size for Longitudinal Response

Interested in comparing two treatments (or exposures), denoted A and B.

Plan to randomize an equal number of subjects, say N, to two groups.

Plan to take n repeated measurements of the response (not necessarily equally spaced measurements).

Two groups to be compared in terms of changes in the mean response over duration of study.

For simplicity, we assume linear trends and define effect of interest as

difference in slopes or rates of change, say δ.

Under null hypothesis of no group difference, i.e., no group $\times$ linear trend interaction, H_0: $\delta = 0$.

Sample size calculation can be simplified so that earlier formula can be used.

This is achieved by considering two-stage model described in Lecture 12.

Stage 1: assume a simple parametric curve (e.g., linear) fits the observed responses for each subject.

Stage 2: individual-specific parameters are then related to covariates that describe the two groups.

Stage 1:

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

where the errors, ϵ_{ij}, are assumed to be independent and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

Stage 2:

Let $\beta_i = (\beta_{1i}, \beta_{2i})'$.

Allow the mean of β_i (i.e., the mean intercept and slope) to depend on group,

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

Note: β_4 is the group difference in the mean slope; $\delta = \beta_4$.

The between-individual variation in the β_i that cannot be explained by group is

$$(\beta_i) = G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix},$$

where $g_{11} = (\beta_{1i})$, $g_{22} = (\beta_{2i})$, and $g_{12} = g_{21} = (\beta_{1i}, \beta_{2i})$.

Recall: Each subject is measured at common set of occasions, $t_1, ..., t_n$.

Let $\widehat{\beta}_{2i}$ denote the ordinary least squares (OLS) estimate of the slope for the i^{th} subject.

Variability of $\widehat{\beta}_{2i}$, say σ^2, is given by

$$\sigma^2 = (\widehat{\beta}_{2i}) = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \; g_{22},$$

where

$$\bar{t} = \frac{1}{n} \sum_{j=1}^{n} t_j.$$

To test if mean slopes are equal in two groups, we can construct the following z-test based on the $\widehat{\beta}_{2i}$:

$$Z = \frac{\overline{\beta}_2^{(A)} - \overline{\beta}_2^{(B)}}{\sigma \sqrt{\frac{1}{N} + \frac{1}{N}}} = \frac{\overline{\beta}_2^{(A)} - \overline{\beta}_2^{(B)}}{\sigma \sqrt{\frac{2}{N}}},$$

where $\overline{\beta}_2^{(A)}$ and $\overline{\beta}_2^{(B)}$ are the sample averages of $\widehat{\beta}_{2i}$ in groups A and B, respectively, and $\sigma^2 = (\widehat{\beta}_{2i})$.

Given estimates of g_{22}, the between-subject variability in slopes, and σ_ϵ^2, the within-subject variability, the sample size can be determined from

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\,\sigma^2}{\delta^2},$$

where now

$$\sigma^2 = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \; g_{22},$$

and δ is group difference in slopes.

Note: This sample size formula is virtually identical to previous formula except σ^2 has two components:
a within-subject variance component, $\sigma_\epsilon^2 \{\sum_{j=1}^{n}(t_j - \bar{t})^2\}^{-1}$, and
a between-subject variance component, $g_{22} = (\beta_{2i})$.

Finally, in a study of length τ, if the n repeated measurements are taken at equally-spaced times $t_1 = 0, t_2 = \tau/(n-1)$, $t_3 = 2\tau/(n-1), ..., t_n = \tau$,

$$\sum_{j=1}^{n} (t_j - \bar{t})^2 = \frac{\tau^2 \, n \, (n+1)}{12 \, (n-1)}.$$

Further examination of this simple formula reveals how sample size (and power) is impacted by:

(i) the length of the study;

(ii) the number of repeated measures; and

(iii) the spacing of the repeated measures.

Note: In general, investigators have little control over the natural heterogeneity of the study population, $g_{22} = (\beta_{2i})$.

Magnitude of σ^2 can be reduced by increasing magnitude of

$$\sum_{j=1}^{n}(t_j - \bar{t})^2.$$

For example, increase length of study or number of repeated measures.

Example

Interested in comparing two treatments (or exposures), denoted A and B.

Plan to randomize an equal number of subjects, say N, to two group.

Plan to take $n = 5$ repeated measurements of the response; 1 at month 0, remainder at 6-month intervals ($\tau = 2$ years).

For simplicity, we assume linear trends and define effect of interest as difference in slopes or rates of change, say δ.

Suppose investigators want to detect minimum $\delta = 1.2$ (e.g., difference in the annual rates of change in the two groups of no less than 1.2).

Based on historical data, investigators posit that between-subject variability in the rate of change, $(\beta_{2i}) \approx 2$ and the within-subject variability, $\sigma_\epsilon^2 \approx 7$.

Finally, investigators desire to have power of 90% when conducting a 2-sided test at the 5% significance level (i.e., $\gamma = 0.1$ and $\alpha = 0.05$).

Given these specifications,

$$\sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} = \frac{12\,(n-1)\,\sigma_\epsilon^2}{\tau^2\, n\,(n+1)} = \frac{12 \times 4 \times 7}{4 \times 5 \times 6} = 2.8,$$

and

$$\sigma^2 = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \ g_{22} = 2.8 + 2.0 = 4.8.$$

The projected N in each group is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2\ 2\sigma^2}{\delta^2} = \frac{(1.96 + 1.282)^2 \times 2 \times 4.8}{1.44} = 70.1.$$

Thus, to ensure power of at least 90% investigators will need to enroll a total of 142 subjects, randomizing an equal number (71) to each group.

Note study of same duration ($\tau = 2$ years) with $n = 3$ repeated measurements, 12 months apart, would require a total of 162 subjects to achieve comparable power.

Alternatively, study over 3 instead of 2 years (and with same retention rate), with $n = 5$ repeated measurements taken 9 months apart, would require a total of 96 subjects to achieve power of at least 90%.

Table 1 displays power as a function of N and n.

Table 1: Power as a function of sample size and the number of equally spaced repeated measurements in a longitudinal study of fixed duration.

	Number of Repeated Measures (n)				
Sample Size (N)	2	4	6	8	10
20	0.37	0.39	0.43	0.47	0.50
40	0.63	0.66	0.72	0.76	0.79
60	0.80	0.83	0.87	0.90	0.93
80	0.90	0.92	0.95	0.97	0.98
100	0.95	0.96	0.98	0.99	0.99

Power when conducting a 2-sided test at the 5% significance level ($\alpha = 0.05$) when $\tau = 2$, $\delta = 1.2$, $(\beta_{2i}) = 2$, and $\sigma_e^2 = 7$.

Longitudinal and Cross-Sectional Information

In certain longitudinal designs, we have cohorts that differ in age measured repeatedly over time.

In such designs, it is possible to estimate the effect of growth or aging from two different sources of information: longitudinal and cross-sectional.

It is possible for these two sources of information to provide conflicting estimates of effects.

For example, when effect of aging is determined from cross-sectional information, it is potentially confounded by cohort effects.

Therefore, important to consider models that allow for separate parameters for the longitudinal and cross-sectional effects.

In slight departure from notation, let t_{ij} denote age of i^{th} subject at j^{th} occasion.

In another departure from notation, let X_{ij} denote vector of time-varying covariates and Z_i denote vector of time-stationary covariates.

$$Y_{ij} = Z_i'\beta_0 + X_{ij}'\beta_1 + e_{ij}$$

$$Y_{ij} = Z_i'\beta_0 + \left(X_{ij}' - X_{i1}' + X_{i1}'\right)\beta_1 + e_{ij}$$

$$Y_{ij} = Z_i'\beta_0 + \left(X_{ij}' - X_{i1}'\right)\beta_1 + X_{i1}'\beta_1 + e_{ij}$$

So this standard model by default assumes the cross-sectional effect of X (represented by X_{i1}) is *equal* to the longitudinal effect (represented by $(X_{ij} - X_{i1})$

The following linear model simultaneously models cross-sectional and longitudinal effects:

$$Y_{ij} = Z_i'\beta_0 + X_{i1}'\beta^{(C)} + (X_{ij}' - X_{i1}')\beta^{(L)} + e_{ij}.$$

This representation allows both cross-sectional effects, $\beta^{(C)}$, and longitudinal effects, $\beta^{(L)}$, to be modelled simultaneously, but separately.

Interpretation of $\beta^{(C)}$ and $\beta^{(L)}$ becomes more transparent when implied models for initial response and subsequent within-subject changes are considered.

First, consider the model for the initial response, Y_{i1},

$$Y_{i1} = Z_i'\beta_0 + X_{i1}'\beta^{(C)} + e_{i1},$$

since $(X_{i1}' - X_{i1}') = 0$.

$\beta^{(C)}$ represents a vector of regression parameters for cross-sectional effects.

Next, consider the model for within-subject changes from the initial response, $Y_{ij} - Y_{i1}$,

$$\begin{aligned}(Y_{ij} - Y_{i1}) &= Z_i'\beta_0 + X_{i1}'\beta^{(C)} + (X_{ij}' - X_{i1}')\beta^{(L)} + e_{ij} \\ &- \ (Z_i'\beta_0 + X_{i1}'\beta^{(C)} + e_{i1}) \\ &= (X_{ij}' - X_{i1}')\beta^{(L)} + (e_{ij} - e_{i1}).\end{aligned}$$

$\beta^{(L)}$ represents a vector of regression parameters for longitudinal effects.

Simple Example

$$Y_{ij} = \beta_0 + \beta_1\text{Gender}_i + \beta^{(C)}\text{Age}_{i1} + \beta^{(L)}(\text{Age}_{ij} - \text{Age}_{i1}) + e_{ij}.$$

First, consider the model for the initial response, Y_{i1},

$$Y_{i1} = \beta_0 + \beta_1\text{Gender}_i + \beta^{(C)}\text{Age}_{i1} + e_{i1},$$

$\beta^{(C)}$ describes how mean response at baseline changes with age at baseline (cross-sectional change).

Next, consider the model for within-subject changes from the initial response, $Y_{ij} - Y_{i1}$,

$$(Y_{ij} - Y_{i1}) = \beta^{(L)}(\text{Age}_{ij} - \text{Age}_{i1}) + (e_{ij} - e_{i1}).$$

$\beta^{(L)}$ describes how within-subject changes in the response are related to within-subject changes in age (longitudinal change).

Formal comparisons can be made by testing H_0: $\beta^{(C)} = \beta^{(L)}$.

Note: When $\beta^{(C)} = \beta^{(L)} = \beta$, the model simplifies to

$$Y_{ij} = Z_i'\beta_0 + X_{ij}'\beta + e_{ij}.$$

In the simple example:

$$Y_{ij} = \beta_0 + \beta_1 \text{Gender}_i + \beta^{(C)}\text{Age}_{i1} + \beta^{(L)}(\text{Age}_{ij} - \text{Age}_{i1}) + e_{ij},$$

the model simplifies to

$$Y_{ij} = \beta_0 + \beta_1 \text{Gender}_i + \beta \text{Age}_{ij} + e_{ij}.$$

However, when $\beta^{(C)} \neq \beta^{(L)}$ but the model does not allow separate estimation of cross-sectional and longitudinal effects,

$$Y_{ij} = Z_i'\beta_0 + X_{ij}'\beta + e_{ij},$$

then β is some weighted combination of $\beta^{(C)}$ and $\beta^{(L)}$ and may not reflect effects of interest.

β confounds the longitudinal effects with the cross-sectional.

Illustration

Suppose three age-cohorts of children, initially aged 5, 6, and 7 years, are measured at baseline and followed annually for three years.

Suppose cross-sectional effect of age on the baseline response is linear, with

$$E(Y_{i1}) = \beta^{(C)} \text{Age}_{i1},$$

(for simplicity, model with intercept=0 is assumed).

Mean response increases linearly with changes in age in each cohort

$$E(Y_{ij} - Y_{i1}) = \beta^{(L)} (\text{Age}_{ij} - \text{Age}_{i1}),$$

but with slope $\beta^{(L)} \neq \beta^{(C)}$.

Figure 1 gives graphical representation of model for mean response versus age, when $\beta^{(C)} = 0.75$ and $\beta^{(L)} = 0.25$.

Note the discernible difference between longitudinal (solid line) and cross-sectional (dotted line) effects of aging.

When these differences are ignored, changes in the mean response (averaged over the three age-cohorts) with age of measurement (dashed line) reflect a combination of $\beta^{(C)}$ and $\beta^{(L)}$.

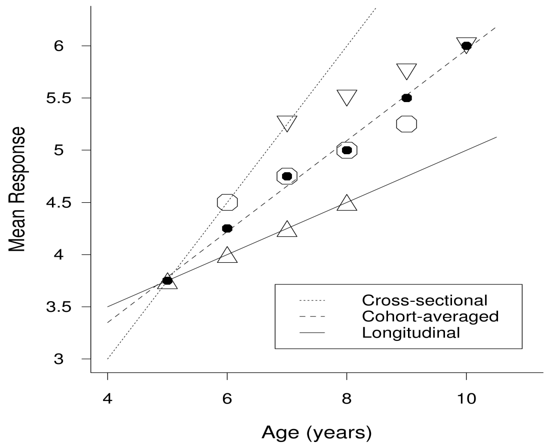

Figure 1: Longitudinal, cross-sectional, and cohort-averaged regression lines for the three age-cohorts: $\triangle$ denotes mean response of children initially aged 5 years; $\bigcirc$ denotes mean response of children initially aged 6 years; and ∇ denotes mean response of children initially aged 7 years. ($\bullet$ denotes averages over cohorts).