Index of Lecture 2–L

Practical Information

Today's lecture: Chapters 1–3 in Diggle et al. (2002),

- introduction to longitudinal (repeated measures) data and analysis,
- discussion of cross-sectional vs. longitudinal designs,
- discussion of consequences of ignoring repeated measures (clustering),
- some simple sample size formulae,
- graphical tools for repeated measures (not in lecture).

General remarks about the book:

- research monography, not a (course) textbook
  $\Rightarrow$ higher level of sophistication,
- more formulae, fewer examples, no problems at the end of chapters,
- standard reference for many methods.

Ignoring Repeated Measures

Consequences of ignoring correlation in longitudinal data:

- incorrect inferences about regression coefficients,
- estimates of regression coefficients which are inefficient (i.e., less precise than possible),
- sub-optimal protection against biases caused by missing data.

Note: estimates may be unbiased even if the correlation is ignored! (bias is not the full story...)

Illustration by simple regression,

$$y_{ij} = \beta_0 + \beta x_j + \varepsilon_{ij} \quad \text{with } x_j = j,$$

where $i \sim$ subjects, $j \sim$ observation within subjects, and the within-subject errors $(\varepsilon_{i1}, \ldots, \varepsilon_{it})$ are correlated so that

$$\begin{aligned} \text{Corr}(\varepsilon_{i1}, \varepsilon_{i2}) &= \rho, \ \ldots, \ \text{Corr}(\varepsilon_{i,t-1}, \varepsilon_{it}) = \rho, \\ \text{Corr}(\varepsilon_{i1}, \varepsilon_{i3}) &= \rho^2, \ \ldots, \ \text{Corr}(\varepsilon_{i,t-2}, \varepsilon_{it}) = \rho^2, \end{aligned}$$

and so on (first-order autoregressive correlation structure). Figure 1.7 in DH compares ordinary least squares (OLS) and "optimal" estimation:

- reported (OLS) SE of $\hat{\beta}_{\text{OLS}}$ agrees with true SE only when ρ very close to 0,
- true SE of $\hat{\beta}_{\text{OLS}}$ slightly larger than SE of $\hat{\beta}_{\text{opt}}$.

Analytical Approaches to Repeated Measures

- summary statistics (response features) — lecture 1L,
- random effects model (or conditional model), e.g.

  $\mathrm{E}(y_{ij}|u_i) = \beta_0 + \beta_1 x_{ij} + u_i$,
  $u_i \sim \mathrm{N}(0, \sigma_u^2)$ — subject random effects,

  $\sim$ model for inference about x's effect for (specific) subjects,

- marginal model, e.g.

  $\mathrm{E}(y_{ij}) = \beta_0 + \beta_1 x_{ij}$,
  additional assumptions about variance/correlation of $(y_{i1}, \ldots, y_{it})$,

  $\sim$ model for inference about x's effect averaged across subjects,

- transitional model, e.g.

  $\mathrm{E}(y_{ij}|y_{i,j-1}) = \beta_0 + \beta_1 x_{ij} + \gamma y_{i,j-1}$,

  $\sim$ model for inference about x's effect on transition (or change) within individuals.

Longitudinal vs. Cross-sectional

Cross-sectional studies (Fig. 1.1(a)) give information on:

- differences between subjects in subpopulations,

Longitudinal studies (Fig. 1.1 (b)+(c)) give information on:

- differences between subjects in subpopulations,
- changes in subjects across conditions (time).

Illustration by simple regression,

$$\begin{aligned} \text{CS}: \; y_{i1} &= \beta_0 + \beta_{\text{CS}} x_{i1} + \varepsilon_{i1}, \\ \text{L}: \; y_{ij} &= \beta_0 + \beta_{\text{CS}} x_{i1} + \beta_{\text{L}}(x_{ij} - x_{i1}) + \varepsilon_{ij}, \end{aligned}$$

where $i \sim$ subject and $j \sim$ observation within subject.

Interpretation:

- $\beta_{\text{CS}} \sim$ expected difference in y across two subpopulations differing by one unit of x,
- $\beta_{\text{L}} \sim$ expected change in y within the same subject for a one unit change in x,
- Fig. 1.1 (c): $\beta_{\text{L}} > 0$, $\beta_{\text{CS}} > 0$,
- Fig. 1.1 (b): $\beta_{\text{L}} < 0$, $\beta_{\text{CS}} > 0$,
- to estimate how individuals change from a CS study we must assume $\beta_{\text{CS}} = \beta_{\text{L}}$,
- (technical): $\beta_{\text{CS}} \neq \beta_{\text{L}} \sim$ contextual effect of x,
- even if $\beta_{\text{L}} = \beta_{\text{CS}}$, estimation is more efficient in longitudinal studies.

Sample Size for Longitudinal Studies

Basic principles for determination of appropriate sample size (required no. of subjects):

- based on either estimation accuracy or statistical power,
- always based on a pre-selected statistical model and effect/statistic,
  - preferably as simple as possible (while incorporating key features), e.g. summary statistic approach and comparison of two groups,
  - desired precision/effect size must be specified, as well as additional model parameters, e.g.
    - in full repeated measures models: within-subject correlation,
    - in normal distribution models: standard dev.,
- simple approximation formulae exist (next slides),
- best (only) general approach is simulation, but not easy.

Software for sample size calculation for clustered data:

- PINT (Tom Snijders, 2-level; `http://stat.gamma.rug.nl/`),
- MPLUS (commercial; `www.statmodel.com/index2.html`),
- simulation approach: any high-level software like R, Stata, SAS...

Sample Size for Slopes, Continuous Outcome

Regression model for comparing groups A vs. B:

$$y_{ij} = \beta_{0,\mathrm{grp}(i)} + \beta_{1,\mathrm{grp}(i)} x_j + \varepsilon_{ij},$$

where $\mathrm{grp}(i) =$ A or B, and $i = 1, \ldots, n$; $j = 1, \ldots, t$.

Sample size based on power for true effect (difference in slopes) $d = \beta_{1A} - \beta_{1B}$,

$$n = \frac{2(z_{1-\alpha} + z_\beta)^2 \sigma^2 (1-\rho)}{\mathrm{SSX}\, d^2},$$

with

- standard dev. σ, sum of squares $\mathrm{SSX} = \sum_j (x_j - \bar{x})^2$,
- one-sided error level α (use $\alpha/2$ for a two-sided test at level α), and power β,
- same within-subject correlations: $\mathrm{Corr}(y_{ij}, y_{ij'}) = \rho$,
- $z_\alpha = \alpha$-percentile of N(0,1), e.g. $z_{0.975} = 1.96$.

Comments and interpretations:

- the required sample size decreases (!) with increasing ρ,
- the required sample size decreases with increasing SSX,
- usual (uncorrected) formula for regression, except for multiplication by factor $(1-\rho)$,
- to avoid approximations: use exact formula/software for regression and multiply by $(1-\rho)$,
- adjustment for other correlation structures possible.

Sample Size for Means

Continuous 2-sample model for comparing groups A vs. B:

$$y_{ij} = \mu_{\mathrm{grp}(i)} + \varepsilon_{ij}, \quad i = 1, \ldots, n; \; j = 1, \ldots, t.$$

Sample size based on power for true effect (mean difference) $d = \mu_A - \mu_B$,

$$n = \frac{2(z_{1-\alpha} + z_\beta)^2 \sigma^2 [1 + (t-1)\rho]}{t\, d^2}.$$

Binary 2-sample model for comparison groups A vs. B:

$$\mathrm{P}(y_{ij} = 1) = p_{\mathrm{grp}(i)}, \quad i = 1, \ldots, n; \; j = 1, \ldots, t\,.$$

Sample size based on power for true effect (probability difference) $d = p_A - p_B$,

$$n = \frac{\left(z_{1-\alpha}\sqrt{2\bar{p}\bar{q}} + z_\beta \sqrt{p_A q_A + p_B q_B}\right)^2 \, [1 + (t-1)\rho]}{t\, d^2},$$

with $q_A = 1 - p_A$, $q_B = 1 - p_B$, $\bar{p} = (p_A + p_B)/2$, $\bar{q} = (q_A + q_B)/2$.

Comments and interpretations:

- usual (uncorrected) formulae (e.g., VER p. 41–42) for total sample size nt, except for multiplication by variance inflation factor $\mathrm{VIF} = [1 + (t-1)\rho]$,
- to avoid approximations: use exact formula/software and multiply by VIF,
- adjustment for other correlation structures possible.