Longitudinal Data Analysis

Adam J Sullivan, PhD

04/16/2018

Longitudinal Data Covariance and Correlation

Ideas for Analyzing Longitudinal Data

- Many of the studies we use attempt to follow people over a period of time rather than just a cross-sectional analysis.
- The primary goal of these longitudinal studies is to characterize the changing in response over time and the factors that influence change.

Ideas for Analyzing Longitudinal Data

- Longitudinal data require different statistical techniques than we have previously considered because
  - Repeated measures on the same individual are often positively correlated.
  - Variability is often heterogeneous across measurement occasions.
- We must consider this correlation and heterogeneity in order to obtain valid inferences.

General Linear Model for Longitudinal Data

- With this model we assume that there are n_i repeated measures on the i$^{\text{th}}$ subject and there is a Y_{ij} observed at each time t_{ij}.
- Also we have various X_{ij}'s that we believe are associated with Y_{ij}.

General Linear Model for Longitudinal Data

- We can consider *linear* regression models for change in mean response over time

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

- The e_{ij} are random error terms with mean 0. We then have that:

$$E\left(Y_{ij} | \mathbf{X}_{ij}\right) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$$

General Linear Model for Longitudinal Data: Vector Format

- Many times we write this in a vector format as:

$$E(Y_{ij}|\mathbf{X}_{ij}) = X_i\beta$$

Assumptions of General Linear Model

1. The individuals represent a random sample from the population.
2. Observations from different individuals are independent, while repeated measurements of the same individual are not assumed to be independent.
3. $Y_{i1}, \ldots, Y_{in_i}$ \$ have a multivariate normal distribution with means

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

   and covariance matrix Σ_i.
4. If there are missing data they are assumed to be "Ignorable".

Missing Data

- We assume that the missing data is ignorable.
- This missingness comes from a subject not having been evaluated at one or more time points.
- In order to be ignorable we need one of the two following situation.

Missing Data

- Before listing these we define the following notation:

$$Y^{(o)} \text{ are observed measurements}$$

$$\text{and}$$

$$Y^{(m)} \text{ are missing measurements}$$

Missing Data Assumptions

1. Data are **Missing Completely at Random** (***MCAR***) when the probability that an individual value will be missing is independent from $Y^{(o)}$ and $Y^{(m)}$. We can then use Maximum Likelihood and other complete cases analysis in order to estimate in these situations.
2. Data are **Missing at Random** (***MAR***) when the probability that an individual will be missing is independent of $Y^{(m)}$ but may be dependent on $Y^{(o)}$. We can then use some likelihood based methods to estimate. This is when a subject's attrition is related to a previous performance.

Modeling Longitudinal Data

- Longitudinal data present two aspects of the data that require modeling:
  1. Mean response over time.
  2. Covariance among the repeated measures.
- We must model both of these jointly.

Mixed Effect Models for Longitudinal Data.

- There are many ways to analyze these models for example:
- Modeling the mean through:
  1. Analysis of Response Profiles.
  2. Parametric or Semi-parametric methods.

Mixed Effects Covariance

- Modeling the covariance through:
  1. Unstructured or arbitrary pattern of covariance
  2. Covariance pattern models.
  3. Random effects covariance structures.

Two-Stage (Two-Level) Formulation

- We will proceed with Linear Mixed effects models.
- They are very useful in longitudinal as well as other hierarchical aspects.
- The basic idea of the model is that we assume
  1. **Stage 1**: A straight line (or more generally a "growth" curve) fits the observed responses for each subject.
  2. **Stage 2**: A Regression model relating the mean of the individual intercepts and slopes to the subject specific effects.

Stage 1

- In the first stage we assume that all subjects have their own unique trajectory.
- So for subject i:

$$Y_{ij} = Z_{ij}\beta_i + \varepsilon_{ij}, \qquad j = 1, \ldots, n_i$$

- where β_i is a vector of subject-specific regression parameters, the errors are typically considered independent within a subject.

16/24

Stage 1: Subject Specific Effects

- Many times we use a model with subject specific intercepts and slope:

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

- So in stage 1 each subject has their own unique regression model.
  - Basically we allow each subject to have their own line.
  - We restrict the covariates in these models to be ones that vary over time.
- Any covariates that do not vary over time or refer to between-subject changes (sex, gender, treatment group, exposure group,…) are not included at this stage.

Stage 2

- In this stage we assume that the β_i's (subject-specific effects) are random and come from some distribution (IE. normal or some other).
- We then model the mean and covariance of the β_i's in the population.

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

Stage 2

- Where
  - A_i are the between subject covariates
  - $b_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix}$ are the random effects for individuals
  - $G = \begin{bmatrix} var(b_{1i}) & cov(b_{1i}, b_{2i}) \\ cov(b_{1i}, b_{2i}) & var(b_{2i}) \end{bmatrix}$ is the covariance matrix for the subject specific effects.

Quick Example

- Consider a treatment vs control setting where we have subject specific intercept, β_{1i}, and slope β_{2i}.
- Then we would model the subject specific effects with a group effect:

$$E(\beta_{1i}) = \beta_1 + \beta_2 \mathrm{GROUP}_i$$
$$E(\beta_{2i}) = \beta_3 + \beta_4 \mathrm{GROUP}_i$$

Quick Example

- Where GROUP_i is an indicator variable for treatment.
- Then in this example we would have the following models for means:

Quick Example

- For the control group:

$$E(\beta_{1i}) = \beta_1$$
$$E(\beta_{2i}) = \beta_3$$

Quick Example

- for the treatment group:

$$E(\beta_{1i}) = \beta_1 + \beta_2$$
$$E(\beta_{2i}) = \beta_3 + \beta_4$$

How do we fit these models:

- One approach has been coined as the "NIH Method" since it was popularized by statisticians working at the NIH.
- What they did was:
  1. Fit a regression to the response data for each subject.
  2. Regress the estimates of the individual intercepts and slopes on subject specific covariates.
- This method was very easy to perform because it did not require any special form of regression software.
- This works very well with balanced data.