Analysis of within-cases normal data[1]

STA442/2101 Fall 2017

[1] See last slide for copyright information.

Overview

Independent Observations

- Most statistical models assume independent observations.
- Sometimes the assumption of independence is unreasonable.
- For example, times series and within cases designs.

Within Cases

- A case contributes a value of the response variable for every value of a categorical explanatory variable.
- As opposed to explanatory variables that are *Between Cases*, in which explanatory variables partition the sample.
- It is natural to expect data from the same case to be correlated, *not* independent.
- For example, the same subject appears in several treatment conditions.
- Hearing study: How does pitch affect our ability to hear faint sounds? Subjects are presented with tones at a variety of different pitch and volume levels (in a random order). They press a key when they think they hear something.
- A study can have both within and between cases factors.

You may hear terms like

- **Longitudinal**: The same variables are measured repeatedly over time. Usually lots of variables, including categorical ones, and large samples. If there's an experimental treatment, its usually once at the beginning, like a surgery. Basically its *tracking* what happens over time.
- **Repeated measures**: Usually, same subjects experience two or more experimental treatments. Usually quantitative explanatory variables and small samples.

A simple model

Think of a problem solving study

- Each case contributes an individual shock that pushes all the data values from that case up or down by the same amount.
- Observations from that case are not independent.
- Example: Matched t-test.

$$\begin{aligned} Y_{i,1} &= \mu_1 + \tau_i + \epsilon_{i,1} \\ Y_{i,2} &= \mu_2 + \tau_i + \epsilon_{i,2} \\ d_i &= (\mu_1 - \mu_2) + (\epsilon_{i,1} - \epsilon_{i,2}) \end{aligned}$$

- The random shock from the case cancels.
- Each case serves as its own control.
- Power is much improved.

Extending the idea

- The random shock is a "random effect."
- The classical normal model approach to repeated measures is based on mixed (fixed and random effects) models.

General Mixed Linear Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

- $\mathbf{X}$ is an $n \times p$ matrix of known constants.
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants.
- $\mathbf{Z}$ is an $n \times q$ matrix of known constants.
- $\mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\Sigma}_b$ unknown but often diagonal.
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.

Random vs. fixed effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\epsilon}$$

- Elements of $\boldsymbol{\beta}$ are called fixed effects.
- Elements of $\mathbf{b}$ are called random effects.
- Models with both are called *mixed*.

Main application of random effects models

A random factor is one in which the values of the factor are a random sample from a populations of values.

- Randomly select 20 fast food outlets, survey customers in each about quality of the fries. Outlet is a random effects factor with 20 values. Amount of salt would be a fixed effects factor.
- Randomly select 10 schools, test students at each school. School is a random effects factor with 10 values.
- Randomly select 15 naturopathic medicines for arthritis (there are quite a few), and then randomly assign arthritis patients to try them. Drug is a random effects factor.
- Randomly select 15 lakes. In each lake, measure how clear the water is at 20 randomly chosen points. Lake is a random effects factor.

One random factor

A nice simple example

- Randomly select 5 farms.
- Randomly select 10 cows from each farm, milk them, and record the amount of milk from each one.
- The one random factor is Farm.
- Total $n = 50$

The idea is that "Farm" is a kind of random shock that pushes all the amounts of milk in a particular farm up or down by the same amount.

Farm is a random shock

$$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij},$$

where

$\mu_.$ is an unknown constant parameter.

$\tau_i \sim N(0, \sigma^2_\tau)$

$\epsilon_{ij} \sim N(0, \sigma^2)$

τ_i and ϵ_{ij} are all independent.

$\sigma^2_\tau \geq 0$ and $\sigma^2 > 0$ are unknown parameters.

$i = 1, \ldots q$ and $j = 1, \ldots, k$

There are $q = 5$ farms and $k = 10$ cows from each farm.

General Mixed Linear Model Notation

$$
\begin{aligned}
Y_{ij} &= \mu_. + \tau_i + \epsilon_{ij} \\
\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}
\end{aligned}
$$

$$
\begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{1,3} \\ \vdots \\ Y_{5,9} \\ Y_{5,10} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} (\mu_.) + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \epsilon_{1,3} \\ \vdots \\ \epsilon_{5,9} \\ \epsilon_{5,10} \end{pmatrix}
$$

Distribution of $Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$

- $Y_{ij} \sim N(\mu_., \sigma_\tau^2 + \sigma^2)$
- $Cov(Y_{ij}, Y_{i,j'}) = \sigma_\tau^2$ for $j \neq j'$
- $Cov(Y_{ij}, Y_{i',j'}) = 0$ for $i \neq i'$
- Observations are not all independent.
- Covariance matrix of $\mathbf{Y}$ is block diagonal: Matrix of matrices.
  - Off-diagonal matrices are all zeros.
  - Matrices on the diagonal $(k \times k)$ have the *compound symmetry* structure

$$\begin{pmatrix} \sigma^2 + \sigma_\tau^2 & \sigma_\tau^2 & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma^2 + \sigma_\tau^2 & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 & \sigma^2 + \sigma_\tau^2 \end{pmatrix}$$

(Except it's 10×10.)

Skipping lots of details

$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$

- Distribution theory.
- Components of variance.
- Testing $H_0 : \sigma^2_\tau = 0$.
- Extension to mixed models.
- Nested effects.
- Choice of F statistics based on expected mean squares.

Repeated measures

Another way to describe *within-cases*

- Sometimes an individual is tested under more than one condition, and contributes a response for each value of a categorical explanatory variable.
- One can view "subject" as just another random effects factor, because subjects supposedly were randomly sampled.
- Subject would be nested within sex, but might cross stimulus intensity.
- This is the classical (old fashioned) way to analyze repeated measures.

Problems with the classical approach

- Normality matters in a serious way for the tests of random effects.
- Sometimes (especially for complicated mixed models) a valid F-test for an effect of interest just doesn't exist.
- When sample sizes are unbalanced, everything falls apart.
  - Mean squares are independent of MSE, but not of one another.
  - Chi-squared variables involve matrix inverses, and variance terms no longer cancel in numerator and denominator.
  - What about covariates? Now it gets really complicated.

A modern approach using the general mixed linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

- $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_b\mathbf{Z}^\top + \sigma^2 I_n)$
- Estimate $\boldsymbol{\beta}$ as usual with $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$.
- Estimate $\boldsymbol{\Sigma}_b$ and σ^2 by maximum likelihood, or by "restricted" maximum likelihood.

Restricted maximum likelihood

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

- Transform $\mathbf{y}$ by the $q \times n$ matrix $\mathbf{K}$.
- The rows of $\mathbf{K}$ are orthoganal to the columns of $\mathbf{X}$, meaning $\mathbf{KX} = \mathbf{0}$.
- Then

$$\begin{aligned} \mathbf{Ky} &= \mathbf{KX}\boldsymbol{\beta} + \mathbf{KZb} + \mathbf{K}\boldsymbol{\epsilon} \\ &= \mathbf{KZb} + \mathbf{K}\boldsymbol{\epsilon} \\ &\sim N(\mathbf{0}, \mathbf{KZ}\boldsymbol{\Sigma}_b\mathbf{Z}^\top\mathbf{K}^\top + \sigma^2\mathbf{KK}^\top) \end{aligned}$$

- Estimate $\boldsymbol{\Sigma}_b$ and σ^2 by maximum likelihood.
- A big theorem says the result does not depend on the choice of $\mathbf{K}$.

Nice results from restricted maximum likelihood

- F statistics that correspond to the classical ones for balanced designs.
- For unbalanced designs, "F statistics" that are actually excellent F approximations — not quite F, but very close.
- R's `nlme4` package and SAS `proc mixed`.

Random Intercept Models

- Drop the complicated classical mixed model machinery.
- Retain the basic good idea.
- Each subject (person, case) contributes an individual shock that pushes all the data values from that person up or down by the same amount. Because cases are randomly sampled (pretend), it's a random shock.
- This is still a mixed model, but it's much simpler.

Example: The Noise study

Females and males carry out a discrimination task under 3 levels of background noise. Each subject contributes a discrimination score at each noise level.

- It's a 2×3 factorial design.
- Sex is between, noise is within.
- Model:

For $i = 1, \ldots, n$ and $j = 1, \ldots, 3$,

$$\begin{aligned} Y_{i,j} &= \beta_0 + \beta_1 s_i + \beta_2 d_{i,j,1} + \beta_3 d_{i,j,2} + \beta_4 s_i d_{i,j,1} + \beta_5 s_i d_{i,j,2} + b_i + \epsilon_{i,j} \\ &= (\beta_0 + b_i) + \beta_1 s_i + \beta_2 d_{i,j,1} + \beta_3 d_{i,j,2} + \beta_4 s_i d_{i,j,1} + \beta_5 s_i d_{i,j,2} + \epsilon_{i,j} \end{aligned}$$

You could say that the intercept is $N(\beta_0, \sigma_b^2)$.

In matrix form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$

For 2 females and 2 males

$$Y_{i,j} = \beta_0 + \beta_1 s_i + \beta_2 d_{i,j,1} + \beta_3 d_{i,j,2} + \beta_4 s_i d_{i,j,1} + \beta_5 s_i d_{i,j,2} + b_i + \epsilon_{i,j}$$

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{41} \\ Y_{42} \\ Y_{43} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \quad \cdots$$

Continuing $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\epsilon}$

$$Y_{i,j} = \beta_0 + \beta_1 s_i + \beta_2 d_{i,j,1} + \beta_3 d_{i,j,2} + \beta_4 s_i d_{i,j,1} + \beta_5 s_i d_{i,j,2} + b_i + \epsilon_{i,j}$$

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{41} \\ Y_{42} \\ Y_{43} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{41} \\ \epsilon_{42} \\ \epsilon_{43} \end{pmatrix}$$

where $cov(\mathbf{b}) = \sigma_b^2 \mathbf{I}_4$

Covariance matrix of $\mathbf{y}$ is block diagonal

With four blocks on the diagonal that look like this:

$$\begin{pmatrix} \sigma^2+\sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma^2+\sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma^2+\sigma_b^2 \end{pmatrix}$$

- This structure is called *compound symmetry*.
- Pause to reflect.
- I like it for lab studies, especially with several responses to the experimental conditions, in a different random order for each subject.
- For longitudinal studies, not so much.
- It implies $Cov(y_t, y_{t+1}) = Cov(y_t, y_{t+100})$.

lme4
Linear Mixed Effects Models

- Download and install the package.
- The `lmer` function acts like an extended version of `lm`.
- We will use just a fraction of its capabilities.

Syntax

`noise1 = lmer(discrim ~ sex*noise + (1 | ident))`

- Response variable $\sim$ Fixed effects + (Random effects)
- `sex*noise` is short for `sex + noise + sex:noise`.
- Specification of fixed effects is like `lm`.
- Specification of random effects looks like $(A|B)$.
  - A is `lm`-like syntax for the random effects. It creates the $\mathbf{Z}$ matrix in $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$.
  - With a new independent copy for every value of B.

Another example

Compare `noise1 = lmer(discrim ~ sex*noise + (1 | ident))`

- Reaction time tested every day for days 0-9 of sleep deprivation.
- Ten observations on each of 18 subjects.
- Roughly linear, and each subject has her own slope and intercept.

$$\texttt{Reaction} \sim \texttt{Days + (Days | Subject)}$$

Random slope and intercept.

Copyright Information