Applied Statistical Methods II

Repeated Measures Model Part II

Single-factor repeated measures model: example

- We have measured the expression of a gene from four different regions of the brain.
- We want to study the regional factor: whether the expressions in different regions are different.
- Let y_{ij} be the expression from subject i, and region j, $i = 1, \ldots, 20$ and $j = 1, \ldots, 4$.
- We have 4 repeated measures within each subject.

The Model from the Text

- $Y_{ij} = \mu_{\cdot\cdot} + \rho_i + \tau_j + \epsilon_{ij}$
- $\sum_j \tau_j = 0$
- $\rho_i \sim$ iid $N(0, \sigma_\rho^2)$
- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- ρ_i and ϵ_{ij} are independent.
- $EY_{ij} = \mu_{\cdot\cdot} + \tau_j$
- $Var(Y_{ij}) = \sigma_\rho^2 + \sigma^2$
- $Cov(Y_{ij}, Y_{ij'}) = \sigma_\rho^2$ for $j \neq j'$
- $Cov(Y_{ij}, Y_{i'j'}) = 0$ for $i \neq i'$
- The correlation coefficient for two observations from the same subject (ICC): $\frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2}$

The Model from the Text (cont.)

- $Y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$
- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- $\rho_i \sim$ iid $N(0, \sigma_\rho^2)$. Why are we assuming $\sigma_\rho^2 \geq 0$?
- If we stack Y_{ij}'s into a vector Y, $\text{Var}\left(Y\right) = \sigma_\rho^2 B + \sigma^2 I_n$. $B \in \mathbb{R}^{n\times n}$ is a **partition matrix**. It partitions samples by individuals:

$$B_{rs} = \begin{cases} 1 & r, s \text{ come from same individual} \\ 0 & \text{otherwise} \end{cases}$$

- Assuming Y_{ij}'s are jointly normal, can you think of a rotation matrix $U \in \mathbb{R}^{n\times n}$ such that the entries of $U^T Y$ are independent? What will the variance of the entries be?
- If $\lambda_{\max}$ is the largest eigenvalue of B, we MUST have $\text{Corr}\left(Y_{ij}, Y_{ij'}\right) = \frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2} \geq \frac{-1}{\lambda_{\max} - 1}$.

The Model from the Text (cont.)

- $Y_{ij} = \mu_{\cdot\cdot} + \rho_i + \tau_j + \epsilon_{ij}$
- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- $\rho_i \sim$ iid $N(0, \sigma_\rho^2)$. Why are we assuming $\sigma_\rho^2 \geq 0$?
- If we stack Y_{ij}'s into a vector Y, $\text{Var}\,(Y) = \sigma_\rho^2 B + \sigma^2 I_n$. $B \in \mathbb{R}^{n \times n}$ is a **partition matrix**. It partitions samples by individuals:

$$B_{rs} = \begin{cases} 1 & r, s \text{ come from same individual} \\ 0 & \text{otherwise} \end{cases}$$

- Assuming Y_{ij}'s are jointly normal, can you think of a rotation matrix $U \in \mathbb{R}^{n \times n}$ such that the entries of $U^T Y$ are independent? What will the variance of the entries be?
- If $\lambda_{\max}$ is the largest eigenvalue of B, we MUST have $\text{Corr}\,(Y_{ij}, Y_{ij'}) = \frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2} \geq \frac{-1}{\lambda_{\max} - 1}$.

The Model from the Text (cont.)

- $Y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$
- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- $\rho_i \sim$ iid $N(0, \sigma_\rho^2)$. Why are we assuming $\sigma_\rho^2 \geq 0$?
- If we stack Y_{ij}'s into a vector Y, $\text{Var}\,(Y) = \sigma_\rho^2 B + \sigma^2 I_n$. $B \in \mathbb{R}^{n \times n}$ is a **partition matrix.** It partitions samples by individuals:

$$B_{rs} = \begin{cases} 1 & r, s \text{ come from same individual} \\ 0 & \text{otherwise} \end{cases}$$

- Assuming Y_{ij}'s are jointly normal, can you think of a rotation matrix $U \in \mathbb{R}^{n \times n}$ such that the entries of $U^T Y$ are independent? What will the variance of the entries be?
- If $\lambda_{\max}$ is the largest eigenvalue of B, we MUST have $\text{Corr}\,(Y_{ij}, Y_{ij'}) = \frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2} \geq \frac{-1}{\lambda_{\max} - 1}$.

The Model from the Text (cont.)

- $Y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$
- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- $\rho_i \sim$ iid $N(0, \sigma_\rho^2)$. Why are we assuming $\sigma_\rho^2 \geq 0$?
- If we stack Y_{ij}'s into a vector Y, $\mathrm{Var}\,(Y) = \sigma_\rho^2 B + \sigma^2 I_n$. $B \in \mathbb{R}^{n \times n}$ is a **partition matrix**. It partitions samples by individuals:

$$B_{rs} = \begin{cases} 1 & r, s \text{ come from same individual} \\ 0 & \text{otherwise} \end{cases}$$

- Assuming Y_{ij}'s are jointly normal, can you think of a rotation matrix $U \in \mathbb{R}^{n \times n}$ such that the entries of $U^T Y$ are independent? What will the variance of the entries be?
- If $\lambda_{\max}$ is the largest eigenvalue of B, we MUST have $\mathrm{Corr}\left(Y_{ij}, Y_{ij'}\right) = \frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2} \geq \frac{-1}{\lambda_{\max} - 1}$.

Sum of squares in mixed effects model

- Assume factor A (ρ) is random and factor B (τ) is fixed
- SS terms are the same as defined in the additive two-way ANOVA, but we drop the index k as $n = 1$, i.e. each pair (i, j) is observed once, for $i = 1, \ldots, a$ and $j = 1, \ldots, b$.
- $SSTO = SSA + SSB + SSE$
  - $SSTO = \sum_{ij} \left(Y_{ij} - \overline{Y_{..}}\right)^2$ has $a \times b - 1$ df.
  - $SSA = b \sum_{i} \left(Y_{i.} - \overline{Y_{..}}\right)^2$ has $a - 1$ df.
  - $SSB = a \sum_{j} \left(Y_{.j} - \overline{Y_{..}}\right)^2$ has $b - 1$ df.
  - $SSE = \sum_{ij} \left(Y_{ij} - \overline{Y_{i.}} - \overline{Y_{.j}} + \overline{Y_{..}}\right)^2$ has $a \times b - a - b + 1$ df.

MS and expectation

- In a fixed two-way ANOVA
  - $E[MSA] = \sigma^2 + b\frac{\sum(\mu_i - \mu_.)^2}{a-1}$
  - $E[MSB] = \sigma^2 + a\frac{\sum(\mu_{.j} - \mu_{..})^2}{b-1}$
  - $E[MSE] = \sigma^2$
- In a mixed effects model
  - $E[MSA] = \sigma^2 + b\sigma_A^2$
  - $E[MSB] = \sigma^2 + a\frac{\sum_j \tau_j^2}{b-1}$
  - $E[MSE] = \sigma^2$
- If the SS term does not involve fixed effect terms, we usually have $SS \sim \frac{E(SS)}{df}\chi^2_{df}$, otherwise it will involve a non-central parameter from the fixed effects.

Testing in mixed effects model using OLS (not MLE)

- Test of the fixed effect B
  - $H_0 : \tau_j = 0$ for all j
  - $MSB/MSE \sim F_{b-1,a(b-1)-b+1}$ under H_0
  - What do you think will happen if we ignore correlation between individuals?
  - If we ignore correlations between individuals:

$$F_* = \frac{E(MSB)}{E(MSE)} \underbrace{=}_{H_0 \text{ true; HW}} \frac{\sigma^2 + a\sigma_A^2}{\sigma^2 + \frac{a(b-1)}{ab-1}\sigma_A^2}$$

    $F_* = 1$ if $\sigma_A^2 = 0$ (no correlation) or $a = 1$ (1 individual).
  - Otherwise, $F_* > 1 \Rightarrow$ anti-conservative inference!!
  - Ignoring correlation is one of the worst, and most common, mistakes in data analysis!
- Test of the random effect A :
  - $H_0 : \sigma_A^2 = 0$
  - $MSA/MSE \sim F_{a-1,ab-a-b+1}$ under H_0

Testing in mixed effects model using OLS (not MLE)

- Test of the fixed effect B
  - $H_0 : \tau_j = 0$ for all j
  - $MSB/MSE \sim F_{b-1,a(b-1)-b+1}$ under H_0
  - What do you think will happen if we ignore correlation between individuals?
  - If we ignore correlations between individuals:

$$F_* = \frac{E(MSB)}{E(MSE)} \underbrace{=}_{H_0 \text{ true; HW}} \frac{\sigma^2 + a\sigma_A^2}{\sigma^2 + \frac{a(b-1)}{ab-1}\sigma_A^2}$$

    $F_* = 1$ if $\sigma_A^2 = 0$ (no correlation) or $a = 1$ (1 individual).
  - Otherwise, $F_* > 1 \Rightarrow$ anti-conservative inference!!
  - Ignoring correlation is one of the worst, and most common, mistakes in data analysis!
- Test of the random effect A :
  - $H_0 : \sigma_A^2 = 0$
  - $MSA/MSE \sim F_{a-1,ab-a-b+1}$ under H_0

Testing in mixed effects model using OLS (not MLE)

- Test of the fixed effect B
  - $H_0 : \tau_j = 0$ for all j
  - $MSB/MSE \sim F_{b-1,a(b-1)-b+1}$ under H_0
  - What do you think will happen if we ignore correlation between individuals?
  - If we ignore correlations between individuals:

$$F_* = \frac{E(MSB)}{E(MSE)} \underbrace{=}_{H_0 \text{ true; HW}} \frac{\sigma^2 + a\sigma_A^2}{\sigma^2 + \frac{a(b-1)}{ab-1}\sigma_A^2}$$

    $F_* = 1$ if $\sigma_A^2 = 0$ (no correlation) or $a = 1$ (1 individual).
  - Otherwise, $F_* > 1 \Rightarrow$ anti-conservative inference!!
  - Ignoring correlation is one of the worst, and most common, mistakes in data analysis!
- Test of the random effect A :
  - $H_0 : \sigma_A^2 = 0$
  - $MSA/MSE \sim F_{a-1,ab-a-b+1}$ under H_0

Testing in mixed effects model using OLS (not MLE)

- Test of the fixed effect B
  - $H_0 : \tau_j = 0$ for all j
  - $MSB/MSE \sim F_{b-1,a(b-1)-b+1}$ under H_0
  - What do you think will happen if we ignore correlation between individuals?
  - If we ignore correlations between individuals:

$$F_* = \frac{E(MSB)}{E(MSE)} \underbrace{=}_{H_0 \text{ true; HW}} \frac{\sigma^2 + a\sigma_A^2}{\sigma^2 + \frac{a(b-1)}{ab-1}\sigma_A^2}$$

    $F_* = 1$ if $\sigma_A^2 = 0$ (no correlation) or $a = 1$ (1 individual).
  - Otherwise, $F_* > 1 \Rightarrow$ anti-conservative inference!!
  - Ignoring correlation is one of the worst, and most common, mistakes in data analysis!
- Test of the random effect A :
  - $H_0 : \sigma_A^2 = 0$
  - $MSA/MSE \sim F_{a-1,ab-a-b+1}$ under H_0

Two-factor repeated measures model

- We have measured the expression of a gene from four different regions of the brain.
- We have two groups of subjects: patients (20) and control (20)
- We want to study
  - the regional factor: whether the expressions in different regions are different.
  - the group factor: whether the expressions in two groups are different.
  - the interaction: whether the regional effect is different in two groups.
- Let y_{ijk} be the expression from subject i, group j and region k.
- Note that subject i is nested in group j.
- Can we treat both subject and group as fixed effects and perform inference on group?

The Model from the Text

- $Y_{ijk} = \mu_{...} + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$, $i = 1, \ldots, s$, $j = 1, \ldots, a$, and $k = 1, \ldots, b$.
- $\sum_j \alpha_j = 0$, $\sum_k \beta_k = 0$, $\sum_j((\alpha\beta)_{jk}) = \sum_k((\alpha\beta)_{jk}) = 0$
- $\rho_{i(j)} \sim$ iid $N(0, \sigma_\rho^2)$, $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$, and ρ_i and ϵ_{ij} are independent. (notation: $i(j) :=$ individual i from group j)
- $EY_{ijk} = \mu_{..} + \alpha_j + \beta_k + (\alpha\beta)_{jk}$
- $Var(Y_{ijk}) = \sigma_\rho^2 + \sigma^2$
- $Cov(Y_{ijk}, Y_{ijk'}) = \sigma_\rho^2$ for $k \neq k'$
- $Cov(Y_{ijk}, Y_{i'j'k'}) = 0$ for $i \neq i'$
- The correlation coefficient for two observations from the same subject (ICC): $\frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2}$
- Again stacking Y_{ijk}'s into Y given us

$$E(Y) = X\gamma, \quad \mathrm{Var}(Y) = \sigma_\rho^2 B + \sigma^2 I_n$$

$X \in \mathbb{R}^{n \times 5}$ the design matrix for fixed effects, γ contains fixed effects, B partitions samples by individuals.

The Model from the Text

- $Y_{ijk} = \mu_{...} + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$, $i = 1, \ldots, s$, $j = 1, \ldots, a$, and $k = 1, \ldots, b$.
- $\sum_j \alpha_j = 0$, $\sum_k \beta_k = 0$, $\sum_j((\alpha\beta)_{jk}) = \sum_k((\alpha\beta)_{jk}) = 0$
- $\rho_{i(j)} \sim$ iid $N(0, \sigma_\rho^2)$, $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$, and ρ_i and ϵ_{ij} are independent. (notation: $i(j) :=$ individual i from group j)
- $EY_{ijk} = \mu_{..} + \alpha_j + \beta_k + (\alpha\beta)_{jk}$
- $Var(Y_{ijk}) = \sigma_\rho^2 + \sigma^2$
- $Cov(Y_{ijk}, Y_{ijk'}) = \sigma_\rho^2$ for $k \neq k'$
- $Cov(Y_{ijk}, Y_{i'j'k'}) = 0$ for $i \neq i'$
- The correlation coefficient for two observations from the same subject (ICC): $\frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2}$
- Again stacking Y_{ijk}'s into Y given us

$$E(Y) = X\gamma, \quad \mathrm{Var}(Y) = \sigma_\rho^2 B + \sigma^2 I_n$$

$X \in \mathbb{R}^{n\times 5}$ the design matrix for fixed effects, γ contains fixed effects, B partitions samples by individuals.

SS and MS terms

- Table 27.5 and Table 27.6
- Note that the subject effect is nested in the group effect.
- Test factor A (group): $\frac{MSA}{MSS(A)} \sim F_{a-1,a(s-1)}$.
- Test factor B (region): $\frac{MSB}{MSB.S(A)} \sim F_{b-1,a(s-1)(b-1)}$.
- Test factor AB (interaction): $\frac{MSAB}{MSB.S(A)} \sim F_{(a-1)(b-1),a(s-1)(b-1)}$.

Summary

- There are many different settings of mixed effects models, we talked about two popular mixed effect ANOVA models under the setting of repeated measures.
- For other different designs, SS and MS terms can be defined analogously.
- $E(MS)$ can be computed and F-tests can be derived.
- For unbalanced designs (or missing values), we will use MLE (rMLE) to estimate.
- MLE framework also works for balanced design, and it offers a unified approach regardless of the complex design.
- In practice, we might include other continuous covariates, and we might have linear fixed effects and random effects, which is beyond ANOVA.

A simple longitudinal data example

- We have measured the expression of a gene over four time points.
- We want to study the time effect.
- Let y_{ij} be the expression from subject i, and time j, $i = 1, \ldots, 20$ and $j = 1, \ldots, 4$.
- We have 4 repeated measures within each subject.

Linear mixed effect model

i is indexes individual, j indexes time.

- $Y_{ij} = \mu_{\cdot\cdot} + \rho_i + \beta T_j + \epsilon_{ij}$
- $\rho_i \sim$ iid $N(0, \sigma_\rho^2)$
- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- ρ_i and ϵ_{ij} are independent.
- Do you like this model?
- $EY_{ij} = \mu_{\cdot\cdot} + \beta T_j$
- $Var(Y_{ij}) = \sigma_\rho^2 + \sigma^2$
- $Cov(Y_{ij}, Y_{ij'}) = \sigma_\rho^2$
- $Cov(Y_{ij}, Y_{i'j'}) = 0.$
- Stacking Y_{ij}'s: $E(Y) = X\gamma$, $\mathrm{Var}(Y) = \sigma_\rho^2 B + \sigma^2 I_n$. $\gamma = (1, \beta)^T$.
- This can be fit by likelihood methods. The analysis of longitudinal data is a big topic. Other courses address this in detail.

Linear mixed effect model

i is indexes individual, j indexes time.

- $Y_{ij} = \mu_{\cdot\cdot} + \rho_i + \beta T_j + \epsilon_{ij}$
- $\rho_i \sim$ iid $N(0, \sigma_\rho^2)$
- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- ρ_i and ϵ_{ij} are independent.
- Do you like this model?
- $EY_{ij} = \mu_{\cdot\cdot} + \beta T_j$
- $Var(Y_{ij}) = \sigma_\rho^2 + \sigma^2$
- $Cov(Y_{ij}, Y_{ij'}) = \sigma_\rho^2$
- $Cov(Y_{ij}, Y_{i'j'}) = 0$.
- Stacking Y_{ij}'s: $E(Y) = X\gamma$, $\text{Var}(Y) = \sigma_\rho^2 B + \sigma^2 I_n$. $\gamma = (1, \beta)^T$.
- This can be fit by likelihood methods. The analysis of longitudinal data is a big topic. Other courses address this in detail.

Fitting general random effect models

Assume $Y \in \mathbb{R}^n$ and let $X \in \mathbb{R}^{n \times p}$ be a design matrix. Suppose

$$E(Y) = X\beta, \quad \mathrm{Var}(Y) = \sum_{s=1}^{b} v_s B_s$$

- In previous examples:
  - $b = 2$, $B_1 = I_n$ and $B_2 = B$, where B partitioned samples by individuals.
  - $v_1, v_2 \geq 0$, $ICC = \frac{v_2}{v_1 + v_2}$.
- Above model is quite general and is applicable to many data types. (We will consider other models (e.g. Gaussian processes) later on).
- For general X, previous work with ANOVA is useless.
- Question: how do we fit this??