Mixed Models

Accounting for clustered/correlated data

ERHS 732

Assumptions for linear models

- For a simple linear model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; or \;\; E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

Assumptions for linear models

- For a simple linear model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; or \;\; E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- We assume

Assumptions for linear models

- For a simple linear model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ \ or \ \ E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- We assume
  - Linearity: $E[Y|X]$ is a linear function of X
  - Normality: $\epsilon_i \sim N(0, \sigma^2)$

Assumptions for linear models

- For a simple linear model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; or \;\; E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- We assume
  - Linearity: $E[Y|X]$ is a linear function of X
  - Normality: $\epsilon_i \sim N(0, \sigma^2)$
  - Homoschedasticity: the Var(ϵ_i) is constant, σ^2

Assumptions for linear models

- For a simple linear model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; or \;\; E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- We assume
  - Linearity: $E[Y|X]$ is a linear function of X
  - Normality: $\epsilon_i \sim N(0, \sigma^2)$
  - Homoschedasticity: the Var(ϵ_i) is constant, σ^2
  - **Independence**: The Y_i's are independent random variables

Correlated data

- When we have multiple observations that are expected to be correlated then the independence assumption fails

Correlated data

- When we have multiple observations that are expected to be correlated then the independence assumption fails
  - These observations could be multiple observations per participant over time

Correlated data

- When we have multiple observations that are expected to be correlated then the independence assumption fails
  - These observations could be multiple observations per participant over time
  - Multiple patients of the same doctor

Correlated data

- When we have multiple observations that are expected to be correlated then the independence assumption fails
    - These observations could be multiple observations per participant over time
    - Multiple patients of the same doctor
    - Multiple residents of the same area

Correlated data

- When we have multiple observations that are expected to be correlated then the independence assumption fails
  - These observations could be multiple observations per participant over time
  - Multiple patients of the same doctor
  - Multiple residents of the same area
- In general if the data are 'clustered' in any way we expect them to be correlated

Correlated data

- When we have multiple observations that are expected to be correlated then the independence assumption fails
  - These observations could be multiple observations per participant over time
  - Multiple patients of the same doctor
  - Multiple residents of the same area
- In general if the data are ‘clustered’ in any way we expect them to be correlated
- This correlation has to be accounted for in the model by assuming a certain covariance structure

Correlated data

- Assume we have longitudinal data with N participants and n observations per person ($N \times n$ total person-time observations)

Correlated data

- Assume we have longitudinal data with N participants and n observations per person ($N \times n$ total person-time observations)
  - Each observation is now Y_{it} with $i = 1, \cdots, N$, and $t = 1, \cdots, n$ and the mean of them is $E[Y_{it}]$

Correlated data

- Assume we have longitudinal data with N participants and n observations per person ($N \times n$ total person-time observations)
    - Each observation is now Y_{it} with $i = 1, \cdots, N$, and $t = 1, \cdots, n$ and the mean of them is $E[Y_{it}]$
- The variance of Y_{it} is a measure of the spread of the values around the mean

Correlated data

- Assume we have longitudinal data with N participants and n observations per person ($N \times n$ total person-time observations)
  - Each observation is now Y_{it} with $i = 1, \cdots, N$, and $t = 1, \cdots, n$ and the mean of them is $E[Y_{it}]$
- The variance of Y_{it} is a measure of the spread of the values around the mean
- The *covariance* is a measure of linear dependence between two random variables (say between Y_{i1} and Y_{i2}, σ_{12})

Correlated data

- Assume we have longitudinal data with N participants and n observations per person ($N \times n$ total person-time observations)
  - Each observation is now Y_{it} with $i = 1, \cdots, N$, and $t = 1, \cdots, n$ and the mean of them is $E[Y_{it}]$
- The variance of Y_{it} is a measure of the spread of the values around the mean
- The *covariance* is a measure of linear dependence between two random variables (say between Y_{i1} and Y_{i2}, σ_{12})
- In a traditional linear regression model we assume that any two observations are independent and the covariance is zero. However with correlated data $\sigma_{12} \neq 0$

Sources of variability

- When we have correlated data the total variance is the sum of different sources of variation

Sources of variability

- When we have correlated data the total variance is the sum of different sources of variation
  - Variability between groups/clusters

Sources of variability

- When we have correlated data the total variance is the sum of different sources of variation
  - Variability between groups/clusters
  - Variability within groups/clusters

Sources of variability

- When we have correlated data the total variance is the sum of different sources of variation
  - Variability between groups/clusters
  - Variability within groups/clusters
- Mixed models separate the types of effects into 'fixed' and 'random', but also quantify the contributions of these types of variation

Mixed models - random intercept

- A simple mixed model is one with fixed effects for covariates X and a random intercept for each group or cluster of data

Mixed models - random intercept

- A simple mixed model is one with fixed effects for covariates X and a random intercept for each group or cluster of data
- For the example with repeated observations per participant a mixed model would look like this

$$Y_{it} = \beta_0 + \beta_1 X_{ij} + b_i + \epsilon_{it}$$

Mixed models - random intercept

- A simple mixed model is one with fixed effects for covariates X and a random intercept for each group or cluster of data
- For the example with repeated observations per participant a mixed model would look like this

$$Y_{it} = \beta_0 + \beta_1 X_{ij} + b_i + \epsilon_{it}$$

- b_i represents the random intercept and we assume that $b_i \sim N(0, \sigma_b^2)$. We can rewrite this as

$$Y_{it} = (\beta_0 + b_i) + \beta_1 X_{ij} + \epsilon_{it}$$

Mixed models - random intercept

- A simple mixed model is one with fixed effects for covariates X and a random intercept for each group or cluster of data
- For the example with repeated observations per participant a mixed model would look like this

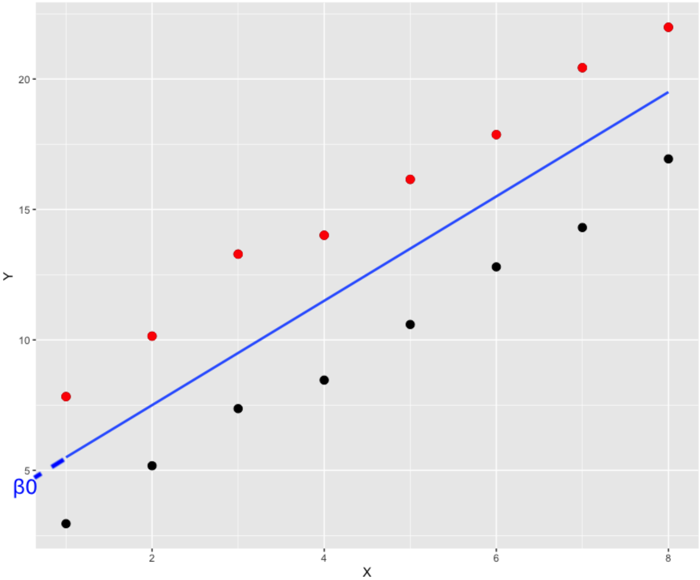
$$Y_{it} = \beta_0 + \beta_1 X_{ij} + b_i + \epsilon_{it}$$

- b_i represents the random intercept and we assume that $b_i \sim N(0, \sigma_b^2)$. We can rewrite this as

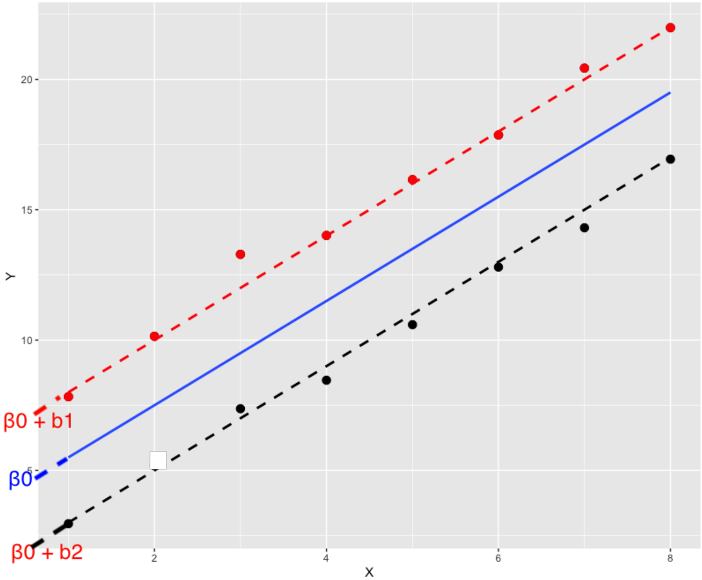
$$Y_{it} = (\beta_0 + b_i) + \beta_1 X_{ij} + \epsilon_{it}$$

- For each participant the intercept is $\beta_0 + b_i$, or in other words the intercept varies randomly around β_0 by a factor of b_i

Mixed models - random intercept



Mixed models - random intercept



Mixed models - random intercept

- We said that $b_i \sim N(0, \sigma_b^2)$ and we still assume $\epsilon_{it} \sim N(0, \sigma^2)$

Mixed models - random intercept

- We said that $b_i \sim N(0, \sigma_b^2)$ and we still assume $\epsilon_{it} \sim N(0, \sigma^2)$
- The total variance is $\sigma_b^2 + \sigma^2$ and σ_b^2 represents between-participant variability while σ^2 is within-participant variability

Mixed models - random intercept

- We said that $b_i \sim N(0, \sigma_b^2)$ and we still assume $\epsilon_{it} \sim N(0, \sigma^2)$
- The total variance is $\sigma_b^2 + \sigma^2$ and σ_b^2 represents between-participant variability while σ^2 is within-participant variability
- σ_b^2 also represents the covariance in this structure and the correlation between two observations from the same participant is $\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$

Mixed models - random slope

- We can also add random slopes for the covariates in the model that vary by group

Mixed models - random slope

- We can also add random slopes for the covariates in the model that vary by group
- For the example with repeated observations per participant a mixed model would look like this

$$Y_{it} = \beta_0 + \beta_1 X_{ij} + b_{0i} + b_{1i} X_{it} + \epsilon_{it}$$

Mixed models - random slope

- We can also add random slopes for the covariates in the model that vary by group
- For the example with repeated observations per participant a mixed model would look like this

$$Y_{it} = \beta_0 + \beta_1 X_{ij} + b_{0i} + b_{1i} X_{it} + \epsilon_{it}$$

- b_{1i} represents the random slope and we assume that $b_{1i} \sim N(0, \sigma^2_{b1})$ on top of the previous assumptions. We can rewrite this as

$$Y_{it} = (\beta_0 + b_i) + (\beta_1 + b_{1i}) X_{ij} + \epsilon_{it}$$

Mixed models - random slope

- We can also add random slopes for the covariates in the model that vary by group
- For the example with repeated observations per participant a mixed model would look like this

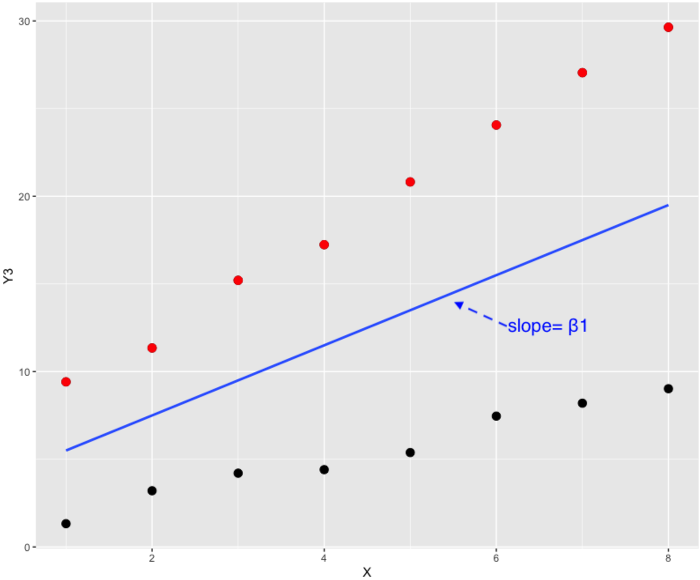
$$Y_{it} = \beta_0 + \beta_1 X_{ij} + b_{0i} + b_{1i} X_{it} + \epsilon_{it}$$

- b_{1i} represents the random slope and we assume that $b_{1i} \sim N(0, \sigma^2_{b1})$ on top of the previous assumptions. We can rewrite this as

$$Y_{it} = (\beta_0 + b_i) + (\beta_1 + b_{1i}) X_{ij} + \epsilon_{it}$$

- For each participant the slope is $\beta_1 + b_{1i}$, or in other words the slope varies randomly around β_1 by a factor of b_{1i}

Mixed models - random intercept



Mixed models - random intercept