Let's talk about $i, j, k, l, m, r, \ldots$

PMHD teaching team

Presentation 1 session

April 18, 2024

DSI
DATA SCIENCE INSTITUTE
UHASSELT
WWW.UHASSELT.BE/DSI

Introduction

A correct model formulation is important for **easy and valid interpretation** of estimation and inferential results

Set the scene – Terminology

- Let Y represent a **population random variable** of interest (i.e., response variable)
- A **sample** $Y_1, Y_2, \ldots, Y_n$ (of size n) is drawn from the study population for estimation and inference (i.i.d.)
- The unit of analysis is denoted here by index i, i.e., referring to **subject** i ($i = 1, \ldots, n$)
- Moreover, consider **covariate information** $\boldsymbol{x}_i$, $i = 1, \ldots, n$ with

$$\boldsymbol{x}_i = [x_{i1} \; x_{i2} \; \ldots \; x_{ip}]^T$$

  a $(p \times 1)$-column vector for subject i
- **Matrix representation**: $\boldsymbol{X}$ is an $(n \times (p+1))$-matrix defined as

$$\boldsymbol{X} = [\mathbf{1} \; \boldsymbol{x}_1 \; \boldsymbol{x}_2 \; \ldots \; \boldsymbol{x}_n]^T = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

Simple linear model formulation

- Consider first a **linear regression model** in which the response variable is regressed against the covariates:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i,$$

  where
  - Y_i represents the outcome of subject $i = 1, \ldots, n$,
  - x_{ij} is the value of the jth covariate related to subject i,
  - $\epsilon_i \sim N(0, \sigma^2)$ are **independent and identically distributed (i.i.d.)** random variables.

- **Model assumptions**:
  - Independence, linearity, homoscedasticity and normality

Simple linear model formulation

- Based on this model, we have
  - $\mu(\boldsymbol{x}_i) := \mathsf{E}(Y_i|\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$
  - Given that $\epsilon_i \sim N(0, \sigma^2)$ are i.i.d., we have that $Y_i|\boldsymbol{x}_i$ are i.i.d. with

$$Y_i|\boldsymbol{x}_i \sim N(\mu(\boldsymbol{x}_i), \sigma^2)$$

  - In **matrix form**:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with

- $\boldsymbol{Y} = [Y_1 \ Y_2 \ \ldots \ Y_n]^T = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \beta_2 \ \ldots \ \beta_p]^T = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$
- $\boldsymbol{\epsilon} = [\epsilon_1 \ \epsilon_2 \ \ldots \ \epsilon_n]^T$

ANOVA model formulation

- Let $x_{i1}, \ldots, x_{ip}$ represent **dummy-variables** related to a single categorical variable (with $p+1$ levels), i.e., for $j = 1, \ldots, p$

$$x_{ij} = \begin{cases} 1 & \text{category for observation } i \text{ is } j \\ 0 & \text{otherwise} \end{cases}$$

- In that case, the linear regression model corresponds to an **ANOVA model** for observations Y^*_{jk} $(k = 1, \ldots, n_j; j = 1, \ldots, p+1)$ which can be formulated as follows:

$$\mu_j := \mathsf{E}(Y^*_{jk}) = \mu + \tau_j,$$

where

  - Y^*_{jk} is the kth observation in group j; $Y^*_{jk} \sim N(\mu_j, \sigma^2)$
  - Sample size n equals sum of group sizes n_j
  - For identifiability reasons, we assume $\tau_{p+1} = 0$
  - Consequently, μ represents the mean of group $p+1$

ANOVA model formulation

- For the aforementioned **parametrization**, we have $\beta_0 = \mu$ and $\beta_j = \tau_j$ for $j = 1, \ldots, p$
- **Can we also consider the parametrization based on the following constraint?**

$$\sum_{j=1}^{p+1} \tau_j = 0$$

Linear mixed model for clustered data

- Consider clustered data with N_c clusters with size n_i for cluster i $(i = 1, \ldots, N_c)$
- Let Y_{ij} represent the jth measurement in cluster i $(j = 1, \ldots, n_i)$
- A **random intercept linear mixed model** can be formulated as follows:

$$Y_{ij} = \beta_0 + b_i + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_p x_{ijp} + \epsilon_{ij},$$

  where
  - ϵ_{ij} i.i.d. random variables with $\epsilon_{ij} \sim N(0, \sigma^2)$
  - $b_i \sim N(0, \sigma_b^2)$ independent

Hierarchical data - more indices (or not)

- Let Y_{ijklm} represent the mth measurement $(m = 1, \ldots, M_l)$ belonging to individual l $(l = 1, \ldots, L_k)$ in household k $(k = 1, \ldots, K_j)$ living in municipality j $(j = 1, \ldots, J_i)$ of region i $(i = 1, \ldots, I)$
- Mean structure (with nested random effects):

$$\mu_{ijklm}(\boldsymbol{b}_{ijkl}) = \beta_0 + \beta_1 x_{ijklm1} + \beta_2 x_{ijklm2} + \ldots + \beta_p x_{ijklmp} + b_{l(k)} + b_{k(j)} + b_{j(i)},$$

  where

$$\boldsymbol{b}_{ijkl} = [b_{l(k)}\ b_{k(j)}\ b_{j(i)}]^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

  with

  - $\boldsymbol{\Sigma}$ the variance-covariance matrix
- Nesting is a property of the data, or rather the experimental design, not the model

Generalized Linear (Mixed) Model – GL(M)M

- Specify **three different components**:
  - Distributional component
  - Systematic component
  - Link function

Tips and tricks

- Do not mix regression and ANOVA notation for covariates, e.g., for observations $i = 1, \ldots, n$, effect garden$_i$ does not make sense (i.e., implies n parameters for garden effect): garden$_i \neq \beta_j$garden$_i$, where garden$_i$ is a dummy-variable
- Definition of indices should be clear (don't reuse indices and make sure indices are defined consistently in model components)
- Order plays a role!
- **Software**: check dummy- versus effect-coding
- Specify (conditional) distributional assumption(s) (also for random effects):

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\lambda_i = \ldots$$

- Check **identifiability issues**: $\beta_0 + \sum_{j=1}^{15} \beta_c C_{ij}$ with C_{ij} compound-specific dummies
- Note: $\log(Y_i) \neq \log(E(Y_i|x_i))$