# Longitudinal data analysis: Assignment 3

**Team B:** **Kendall Brown** *r0773111* **Raïsa Carmen** *s0204278*
**Stefan Velev** *r0924289* **Adhithya Unni Narayanan** *r0776057*

## 1 Introduction

First, the *TIME* variable is rounded to the nearest integer value. Hearing thresholds will be explored, both as a continuous variable and as a trichotomized (ordinal) variable with the following three levels:

- $\leq$ 6 dB: Excellent hearing
- over 6 and $\leq$ 25 dB: Normal hearing
- $\geq$ 26 dB: Hearing loss

### 1.1 Missingness exploration

After discretizing the *TIME* variable, we consider a subject to be missing at a certain time instance if there is no measurement for that subject at that time. It should be noted that, if the subject is not missing (18.16% of TIME-subject instances), we usually (16.88% of TIME-subject instances) have two measurements (one for each ear) at each time instance. In fact, the average number of measurements per subject at each time instance is 0.35, and maximum 4.

Figure 1 was created using the *visdat* package. It shows all subjects, ordered from youngest (in the top) to oldest (in the bottom) and whether or not their data is missing at a certain time instance (on the x-axis). The percentages on top shows the percentage of missingness at each time instance. It is clear from Figure **??** that the missingness is not monotone; subjects may be missing at one time instance and come back later. Since there are too many possible missingness patterns with 23 time instances ($2^{23}$), we do not give an overview of the number of subjects that follow each possible pattern. Instead, figure 2 shows, for each time instance, the number of subjects that:

- are *present*: when the subject's hearing is measured at time $t$ and $t-1$
- are *missing*: when the subject is missing at time $t$ and $t-1$
- *drop out*: when the subject's hearing is measured at time $t-1$ but not at time$t$
- *return*: when the subject's hearing is measured at time $t$ but not at time $t-1$

Figure 2 clearly shows that subject rarely are measured two years in a row, most are not measured at $t = 1$, and the number of subjects that stay missing gradually increases as time passes.

Lastly we explore whether the missingness can be explained by the data by fitting a mixed model to a dataset where $R_it$ is equal to one if the hearing threshold is missing and zero otherwise:

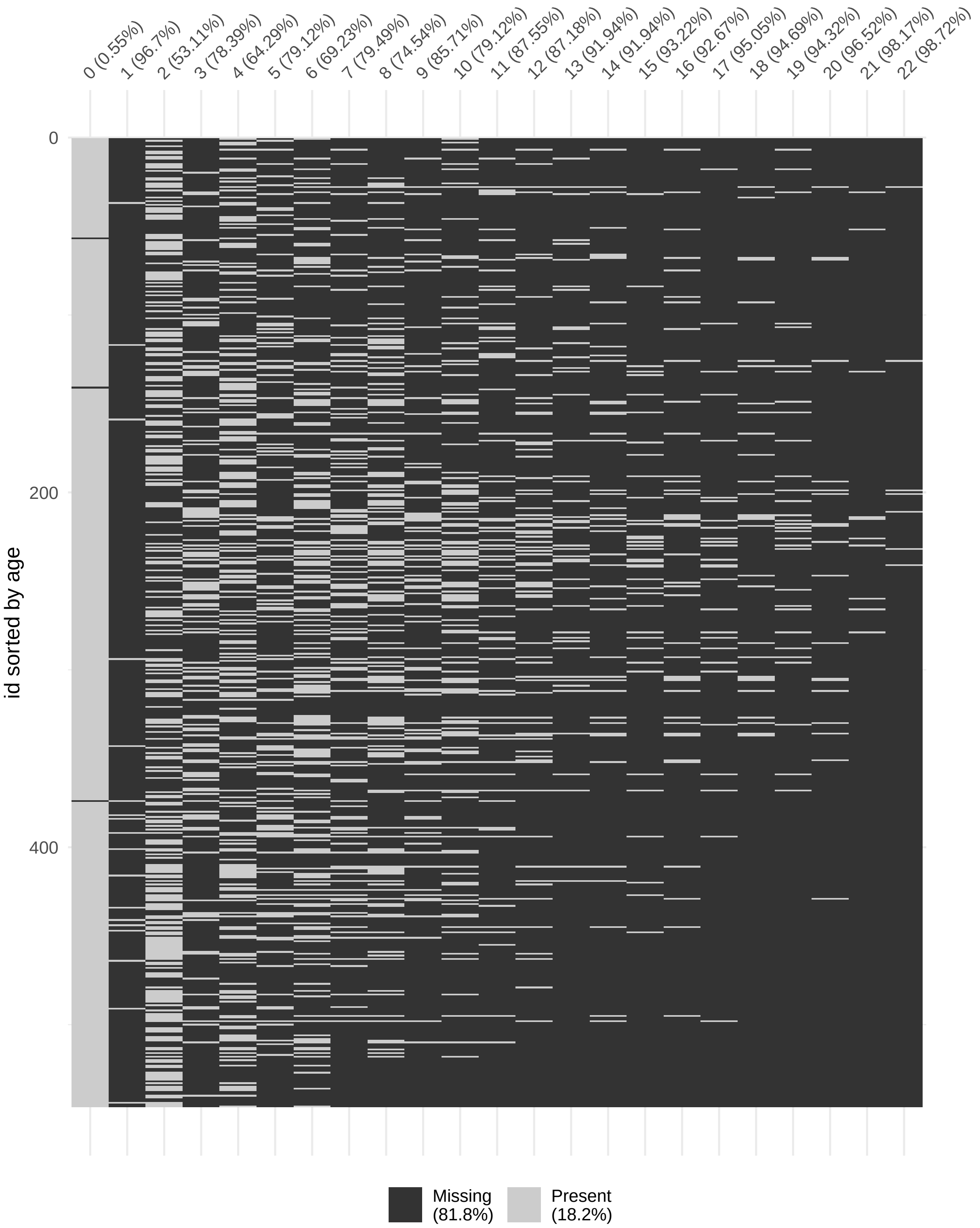

Figure 1: Visual inspection of missingness for different ages at different time instances.

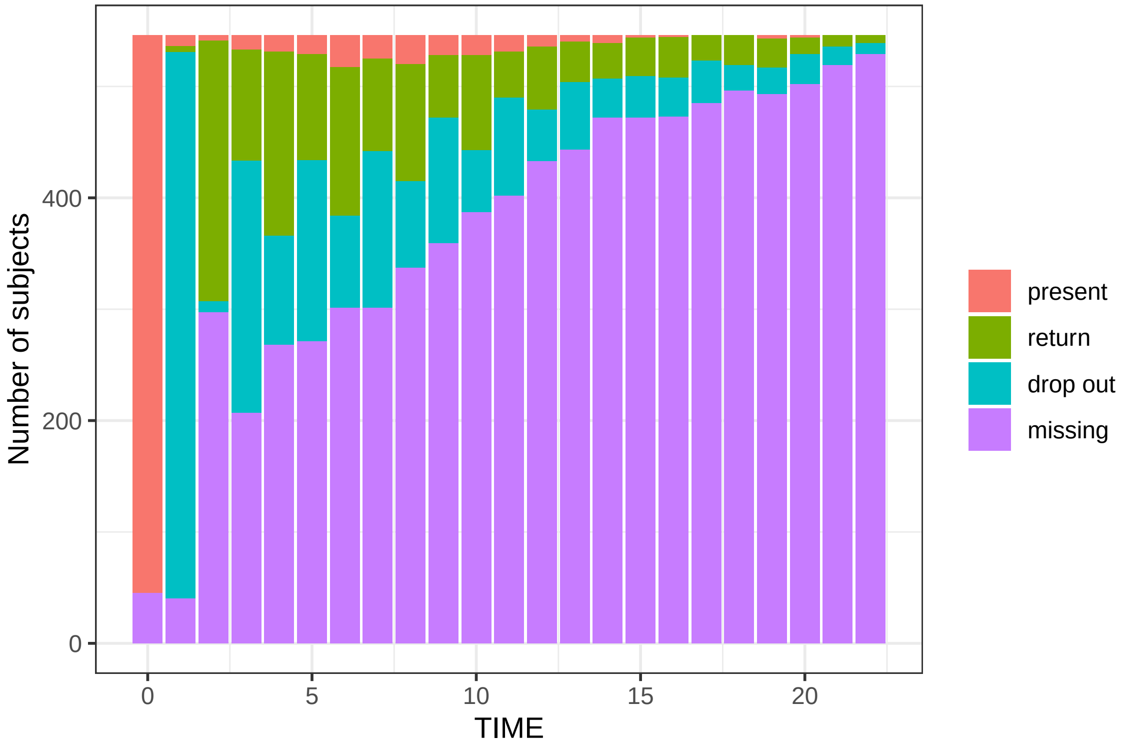

Figure 2: Number of subjects the are present, return, drop out or are missing at each time instance.

$$\begin{cases} logit(R_{it}) = \beta_0 + \beta_1 TIME_{it} + \beta_3 side_{it} + \beta_4 age_{it} + b_i \\ b_i \sim N(0, \sigma^2) \end{cases} \quad (1)$$

The variable *age* was standardized to get convergence in the model. Table @ref(tab:missing_mixedmodel) shows that the only truely significant variable is TIME; as time increases, subjects are more likely to be missing. We can therefore assume missingness at random (MAR).

# 2 Methodology

First, a direct likelihood analysis is compared with multiple imputation in the ?? continuous/discrete??? case. Next, weighted generalized estimating equations are compared with 'multiple-imputation generalized estimating equations'. Lastly, a sensitivity analysis is performed

All analysis was done in R. All scripts are freely available at this git repository.

# 3 Results

## 3.1 Direct likelihood analysis versus multiple imputation

Q4

### 3.2 Weighted generalized estimating equations versus ‘multiple-imputation generalized estimating equations’

Q5

### 3.3 Sensitivity analysis

Q6

## 4 Bibliography