Longitudinal data analysis: Assignment 3

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

1 Introduction

This report assesses the evolution of hearing thresholds over time for a sample of 546 healthy male volunteers. The data originates from the famous Baltimore Longitudinal Study of Aging (BLSA). Previous research showed a change in hearing threshold for all age groups but especially the older population (Brant and Fozard 1990). In this report, we will especially take care of missingness in the data.

First, the *TIME* variable is rounded to the nearest integer value. As such, we aim to balance the dataset with equally-spaced time instances when hearing thresholds are measured.

Hearing thresholds will be explored, both as a continuous variable and as a trichotomized (ordinal) variable with the following three levels:

- ≤ 6 dB: Excellent hearing
- over 6 and ≤ 25 dB: Normal hearing
- ≥ 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 1 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 $t-1$

Table 1: A mixed model to predict missingness.

Variable	Estimate
Intercept	2.22 ***
TIME	0.24 ***
sideright	-0.19 ***
age	0.10
R_{t-1}	-2.67 ***
sigma	2.31

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 $R_i t$ is equal to one if the hearing threshold is missing and zero otherwise:

$$\begin{cases} logit(R_{it}) = \beta_0 + \beta_1 TIME_{it} + \beta_3 side_{it} + \beta_4 age_{it} + \beta_5 R_{it-1} + b_i \\ b_i \sim N(0, \sigma^2) \end{cases} \tag{1}$$

The variable *age* was standardized to get convergence in the model. Table 1 shows that TIME is significant; as time increases, subjects are more likely to be missing. We can therefore assume missingness at random (MAR). Left ear measurements are also more likely to be missing. A subject is also less likely to be missing at time t if he was missing at time $t - 1$. This can be seen especially in the first couple of years in figure 1: all but 3 subjects are measured at $t = 0$, almost no-one is measured at $t = 1$ and many are measured again at $t = 2$.

2 Methodology

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

For imputation, the *mice* library is used (Buuren and Groothuis-Oudshoorn 2011) and different imputation techniques were tested: Predictive mean matching, Bayesian linear regression, Unconditional mean imputation, and imputation by random forests.

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

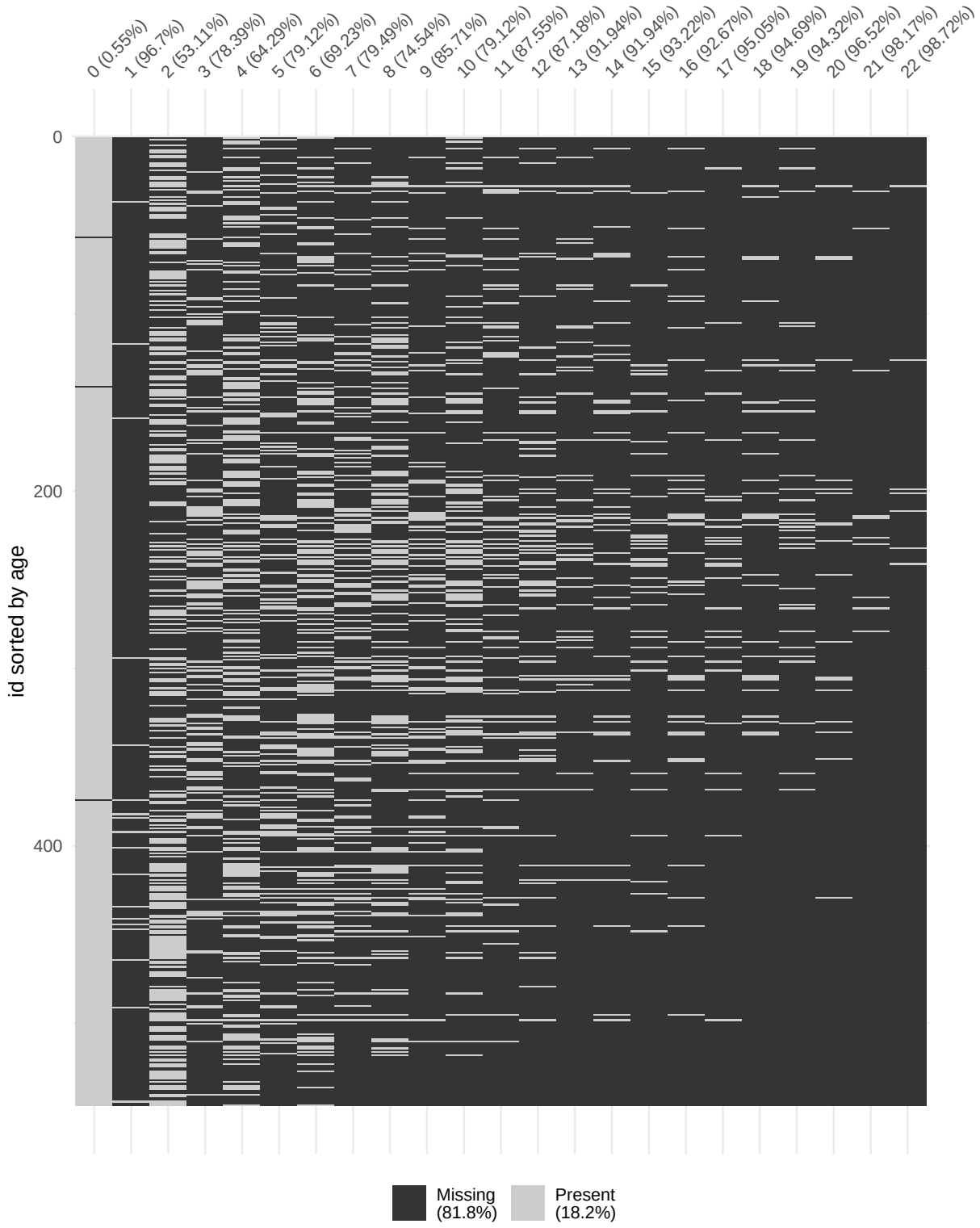

Figure 1: Visual inspection of missingness for different ages at different time instances.

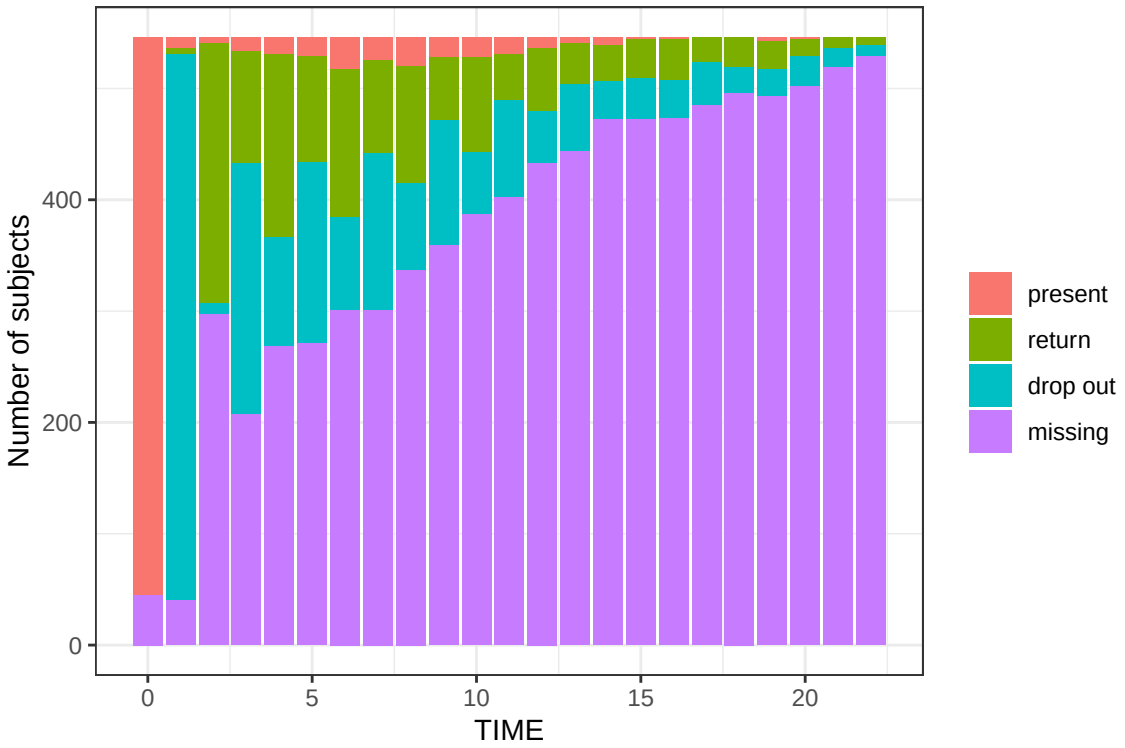

Figure 2: Number of subjects the are present, return, drop out or are missing at each time instance.

3 Results

3.1 Direct Likelihood Analysis

For the purposes of this section we will analyze the hearing data respective to the continuous response. Two popular ways to analyze data sets with missing data is to use either a direct likelihood method or an analysis through multiple imputation. In a direct likelihood analysis, missing values are assumed to be ignorable (Beunckens, Molenberghs, and Kenward 2005). Drafting a linear mixed effects model within R allows for the direct calculation for each patients' respective likelihood functions. The R-package *nlme* allows for this model to be fit in a way such that the parameter estimates are representative of their parameters' maximum log-likelihood. Additionally, we will compare models of differing covariance structures: "simple" and "compound symmetry." Summary statistics for these models can be seen in Table 2.

Parameter estimates from both covariance structures remain stable.

3.1.1 Multiple Imputation

We shall now draft and evaluate models generated by multiple imputation methods. Here, artificial data points are generated according to the algorithms: Predictive mean matching, Bayesian linear regression, Unconditional mean imputation, and imputation by random forests. As we are dealing with continuous data, we also consider two covariance matrix structures: "simple" and "compound symmetry". Being generated from an aggregation of 10 simulations, these sim-

Table 2: Direct likelihood with MI, simple cov. struct.

	Mean imputation	Random forest	Bayesian linear regression	Predictive mean matching
Variable	Estimate (std.error)	Estimate (std.error)	Estimate (std.error)	Estimate (std.error)
(Intercept)	4.294 (0.582)***	-1.449 (1.533)	3.657 (2.806)	4.378 (2.795)
age	0.070 (0.024)**	0.138 (0.064)*	-0.164 (0.119)	-0.190 (0.118)
TIME	0.268 (0.011)***	0.575 (0.025)***	0.319 (0.018)***	0.305 (0.017)***
learning	-0.516 (0.123)***	1.047 (0.243)***	0.884 (0.184)***	0.764 (0.181)***
I(age^2)	0.000 (0.000)	0.000 (0.001)	0.004 (0.001)***	0.004 (0.001)***
age:TIME	-0.005 (0.000)***	-0.009 (0.000)***	0.000 (0.000)	0.001 (0.000)*

Table 3: Direct likelihood with MI, compound symmetry cov. struct.

	Mean imputation	Bayesian linear regression	Predictive mean matching
Variable	Estimate (std.error)	Estimate (std.error)	Estimate (std.error)
(Intercept)	4.294 (0.582)***	3.657 (2.806)	4.378 (2.795)
age	0.070 (0.024)**	-0.164 (0.119)	-0.190 (0.118)
TIME	0.268 (0.011)***	0.319 (0.018)***	0.305 (0.017)***
learning	-0.516 (0.123)***	0.884 (0.184)***	0.764 (0.181)***
I(age^2)	0.000 (0.000)	0.004 (0.001)***	0.004 (0.001)***
age:TIME	-0.005 (0.000)***	0.000 (0.000)	0.001 (0.000)*

ulated data points are then used in conjunction with the real data to draft appropriate models. Their summary statistics are displayed in Table 2 and 3.

Respective to the different covariance structures, parameter estimates are largely the same between model pairs of opposing covariance patterns. For this reason, the following analysis will only consider the models fit according to a simple covariance structure. Comparing model parameter estimates shows that the models do differ in regards to how the fixed effects regression coefficients are computed; with no two models agreeing perfectly on every parameter estimate in regards to sign or magnitude. As these are mixed models, interpreting the fixed effects without consideration of the random effects is innapropraite. Additionally, as data sets are non-standard amongst each of these models, many traditional metrics such as AIC and BIC would also be inapropriate. Instead we shall focus on predicive metrics as these should inform us of each of the models potential practical implications.

Table 4: A comparison of direct likelihood and MI methods

Method	Min SR	Max SR	SR Int. Length
Predictive mean matching	-4.711	7.049	11.761
Bayesian linear regression	-5.041	13.342	18.384
Unconditional mean imputation	-4.518	6.766	11.284
Random forests	-6.766	5.563	12.329
Direct likelihood	-4.039	6.709	10.748

3.1.2 Direct Likelihood vs. Multiple Imputation

To cross evaluate these models and the model drafted from a direct likelihood approach, we shall consider an analysis of the empirically observed within-group standardized residuals. For the models drafted through multiple imputation, to minimize potential bias inflation from the imputation processes, we shall consider a worst case scenario; where these residuals are assumed to rest within the largest possible observed interval between the 10 models (Table 4).

From this metric, we observe that the direct likelihood (DL) method results in the smallest interval of empirically observed within-group standardized residuals, with unconditional mean imputation (Mean) potentially resulting in largest degree of error. As will be discussed in further sections, the impact of predictive bias will be further explored to generate a much more thorough understanding of how well these models may preform.

3.2 Weighted generalized estimating equations versus 'multiple-imputation generalized estimating equations'

In this section we will first explain our specification of the weighted generalized estimating equations, then we will explain the methodology behind the 'multiple-imputation generalized estimating equations' and finally we will compare the models.

3.2.1 Weighted generalized estimating equations

First we specify the weights by using the inverse dropout probability methodology described in the lecture notes. The logistic regression model to predict missingness(R) is the following;

$$\begin{cases} logit(R) = \beta_0 + \beta_k TIME_k + \beta_3 side + \beta_4 age + b \\ b \sim N(0, \sigma^2) \\ k \in 0, 1, \dots, 22 \end{cases} \tag{2}$$

It should be noted that we created a dummy variable for each period in the TIME variable with the first being $TIME = 0$ and the last $TIME = 22$. There are certain periods (e.g. $TIME = 1, 3$) with very few observations. We specified the dummy variables with the hope of capturing any such effects. The model above had the best performance in terms of AIC.

After specifying the weights, we proceed with model selection of the Weighted GEE. For this purpose, we use QIC and a greedy methodology. Starting from a fully specified model we remove redundant variables from the model until we cannot simplify the model any more without sacrificing performance. The final model we chose contains the variables

$$age + TIME + age^2 + age * TIME$$

As a final step we explore the correlation structure of the model using the QICu. The best performing is an independent correlation structure. The model coefficients as well as its error can be seen in the table bellow. We note that the variable age is not significant but we nevertheless include it as age^2 is significant. Further we note that this model is almost the same as the best performing model for the normal un-weighted GEE, the only difference being that the later also includes the learning effect dummy variable for the first observation.

3.2.2 Multiple imputation GEE

In this section, we will discuss the multiple imputation GEE approach. We employ the same 4 different imputation methodsas in section 3.1.1.: Predictive mean matching, Bayesian linear regression, Unconditional mean imputation and imputation by Random Forest. For each, we generate 10 Fully conditional and 10 Monotone multiple imputations leading to a total of 80 imputed data sets (40 FCS and 40 Monotone). The main idea of this method is to replace missing values with M plausible values drawn from the conditional distribution of the missing values given the observed data. This conditional distribution represents how uncertain we are about the right value to impute. We evaluate 3 Multiple imputation models:firstly, a model based on all imputed data sets, both monotone and FCS, secondly a model based only on the FCS imputation, and thirdly a model based only on the monotone imputation.

From each imputed data set, we evaluate $\hat{\beta}^k$ using the same model and correlation structure as above. Here, we note a few important points. First, if we perform a model selection on the imputed data sets, we notice that there are non trivial differences between the best performing model and its correlation structure for the different imputed data sets. We hypothesise that these differences arise from the large number of values we need to impute. For most subjects, we have more imputed data than actual observations. Further, if we do not impute values after the last true measurement of each subject, which greatly reduces the number of imputed values, both models and correlation structure stabilize into models similar to the best performing un-weighted GEE. Secondly, if we impute less data points the performance of the MI GEE on the true data tends to be better. Nevertheless we will conduct the analysis with the full missing data imputation from $TIME = 0$ to $TIME = 22$.
After having evaluated all models based on the imputed data sets we pool the results as follows. The coeficients are evaluated as

$$\overline{\hat{\beta}} = \frac{1}{M} \sum_{k=1}^{M} \hat{\beta}^k$$

Table 5: Model results for weighted GEE and GEE with MI.

	Weighted GEE		MI GEE		MI-Mon GEE		MI-FCS GEE	
	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err
(Intercept)	5.34	2.559	3.496	2.396	2.732	2.715	4.259	1.773
age	-0.173	0.116	-0.078	0.122	-0.033	0.154	-0.124	0.049
TIME	-0.311	0.131	0.24	0.155	0.357	0.124	0.124	0.071
I(age^2)	0.004	0.001	0.003	0.002	0.002	0.002	0.003	0.001
age:TIME	0.009	0.003	-0.001	0.004	-0.003	0.004	0.002	0.002
	RMSE	3.731	RMSE	5.74	RMSE	5.559	RMSE	5.922

Table 6: Model results for different imputation methods.

	MI GEE Mean		MI GEE Norm		MI GEE Pmm		MI GEE Rf	
	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err
(Intercept)	5.373	1.3	3.291	1.249	3.881	1.291	1.438	3.225
age	-0.005	0.08	-0.147	0.049	-0.169	0.053	0.008	0.151
TIME	0.189	0.092	0.215	0.104	0.193	0.118	0.364	0.21
I(age^2)	0.001	0.001	0.004	0	0.004	0.001	0.002	0.002
age:TIME	-0.002	0.003	0.002	0.002	0.002	0.002	-0.004	0.005
	RMSE	3.095	RMSE	6.885	RMSE	6.748	RMSE	6.233

and an estimate of the covariance matrix of $\overline{\hat{\beta}}$ is given by

$$V = W + \left(\frac{M+1}{M}\right) B$$

where

$$W = \frac{1}{M}\sum_{k=1}^{M} U^k \text{ and } B = \frac{1}{M-1}\sum_{k=1}^{M}\left(\hat{\beta}^k - \overline{\hat{\beta}}\right)\left(\hat{\beta}^k - \overline{\hat{\beta}}\right)'$$

here, W measures the within-imputation variability and B measures the between-imputation variability. The between-imputation variability is much larger, and it decreases if we reduce the number of imputed values.

In Table 5 we can see the coefficients for the weighted GEE, the Multiple Imputation GEE based on both monotone and FCS imputation, the MI GEE based only on monotone imputation and finally the MI GEE based only on FCS imputation. At the bottom, we also record the Root Mean Squared Error for each model. We observe that the weighted GEE outperforms all 3 of the MI GEE models. We hypothesize that the reason for this is the large number of values which we needed to impute. Within the MI GEE the one based on monotone imputation performs the best.

Next we will take a look at the performance of the different imputation methods individually. In

Table 7: Model results for different imputation methods with monotone imputation.

	MI GEE Mean		MI GEE Norm		MI GEE Pmm		MI GEE Rf	
	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err
(Intercept)	4.21	0.232	3.728	1.231	4.55	1.181	-1.561	0.739
age	0.07	0.009	-0.159	0.05	-0.192	0.053	0.15	0.025
TIME	0.273	0.011	0.301	0.029	0.295	0.041	0.558	0.043
I(age^2)	0	0	0.004	0	0.004	0.001	0	0
age:TIME	-0.005	0	0	0.001	0.001	0.001	-0.009	0.001
	RMSE	2.051	RMSE	7.108	RMSE	7.013	RMSE	6.063

Table 8: Model results for different imputation methods with FCS imputation.

	MI GEE Mean		MI GEE Norm		MI GEE Pmm		MI GEE Rf	
	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err
(Intercept)	6.537	0.664	2.854	1.144	3.211	1.036	4.436	1.063
age	-0.08	0.028	-0.136	0.046	-0.147	0.043	-0.135	0.045
TIME	0.106	0.042	0.128	0.07	0.092	0.064	0.17	0.079
I(age^2)	0.002	0	0.004	0	0.004	0	0.003	0
age:TIME	0	0.001	0.003	0.001	0.004	0.001	0.001	0.002
	RMSE	4.14	RMSE	6.662	RMSE	6.484	RMSE	6.403

Table 6 we discern between imputation methods but not between monotone or FCS; we average across the 10 monotone AND 10 FCS imputations. We see that the Unconditional mean imputation outperforms the other types of imputation with Random Forest imputation after it with nearly twice as high RMSE. The MI Method with unconditional mean imputation considering both monotone and FCS outperforms the weighted GEE. This gives rise to the question whether or not there is a difference between monotone and FCS imputation between different methods. We can see the results below.

In Table 7 and 8, we compare the monotone and FCS imputation across all 4 different imputation methods. By far the best performing is the unconditional mean monotone imputation. It outperforms every other methodology and is the reason why the averaged unconditional mean method outperformed the weighted GEE.

3.3 Sensitivity analysis

In the previous sections, some sensitivity analysis was done with respect to the different possible imputation methods. Usually, the model results were quite stable. However, all of the previous models assume MAR. In this section, we explore whether the results are robust if the MAR as-

sumption is not realistic and the missingness is not at random (MNAR). Whether or not there is MAR or MNAR cannot be easily and directly tested since we can never know for sure what the hearing threshold would have been if it is missing (Verbeke et al. 2001; Van Steen et al. 2001). It is likely that different assumption on the missingness mechanism will lead to different conclusions, especially if the dropout is higher than 10% (Li et al. 2017). The dropout/missingness is much higher in our dataset and, results are thus expected to be highly impacted.

We use imputation under a MNAR mechanism by the NARFCS method (see Tompsett et al. 2018 and [this useful vignette]{https://github.com/moreno-betancur/NARFCS/blob/master/Vignette.md}). This method allows us the test the effect of different shifts in the MNAR imputed values on the model coefficients. These shifts may also depend on other covariates (such as age and time, see the unidentifiable part in the equation below). Notice that the unidentifiable part will only come into play when an observation is missing ($R_y = 1$). More concretely, The imputed values are of the following form:

$$E(Y|X, R_Y) = \underbrace{\beta_{Y0} + \beta_{Yage}age + \beta_{Yage2}age^2 + \beta_{Ytime}TIME + \beta_{Ytimeage}TIME : age}_{\text{IDENTIFIABLE PART}} \tag{3}$$

$$+ \underbrace{R_Y(\delta_{Y0} + \delta_{Yage}age + \delta_{Yage2}age^2 + \delta_{Ytime}TIME + \delta_{Ytimeage}TIME : age)}_{\text{UNIDENTIFIABLE PART}}. \tag{4}$$

In practice, it is suggested that the sensitivity parameters (δ's) are discussed and chosen with practitioners. For instance, in the hypothetical case that people were asked not let their hearing be measured when they have a cold, the hearing threshold might be higher if the data is missing and practitioners may have some intuition as to how much higher. Here, we don't have any information from practitioners and implement the following shifts as an illustration:

- A shift in the intercept: $\delta_{Y0} \in \{-5; -2; 0; 2; 5\}$. This means that the imputation for the missing values will differ between -5 to 5 dB from expected values in MAR, on average.
- A shift in the effect of age $\delta_{Yage} \in \{-1; -0.5; 0; 0.5; 1\}$.
- A shift in the effect of time $\delta_{Ytime} \in \{-1; -0.5; 0; 0.5; 1\}$.

In each of the 125 (555) scenarios, We will impute the data ten times and will fit a GEE model on the imputed data. Final results in each scenario are then obtained using the same methodology as in section 3.2.2.

```
## Warning: Using alpha for a discrete variable is not advised.
```

```
## Warning: Using alpha for a discrete variable is not advised.
```

Figure 3 shows the estimated intercepts and 4 shows all other parameter estimates.

A shift in the intercept for imputed data (δ_{Y0}) of 1 dB results in an almost equal shift in the estimated intercept from the gee model (0.92 dB). This was expected because such a large part

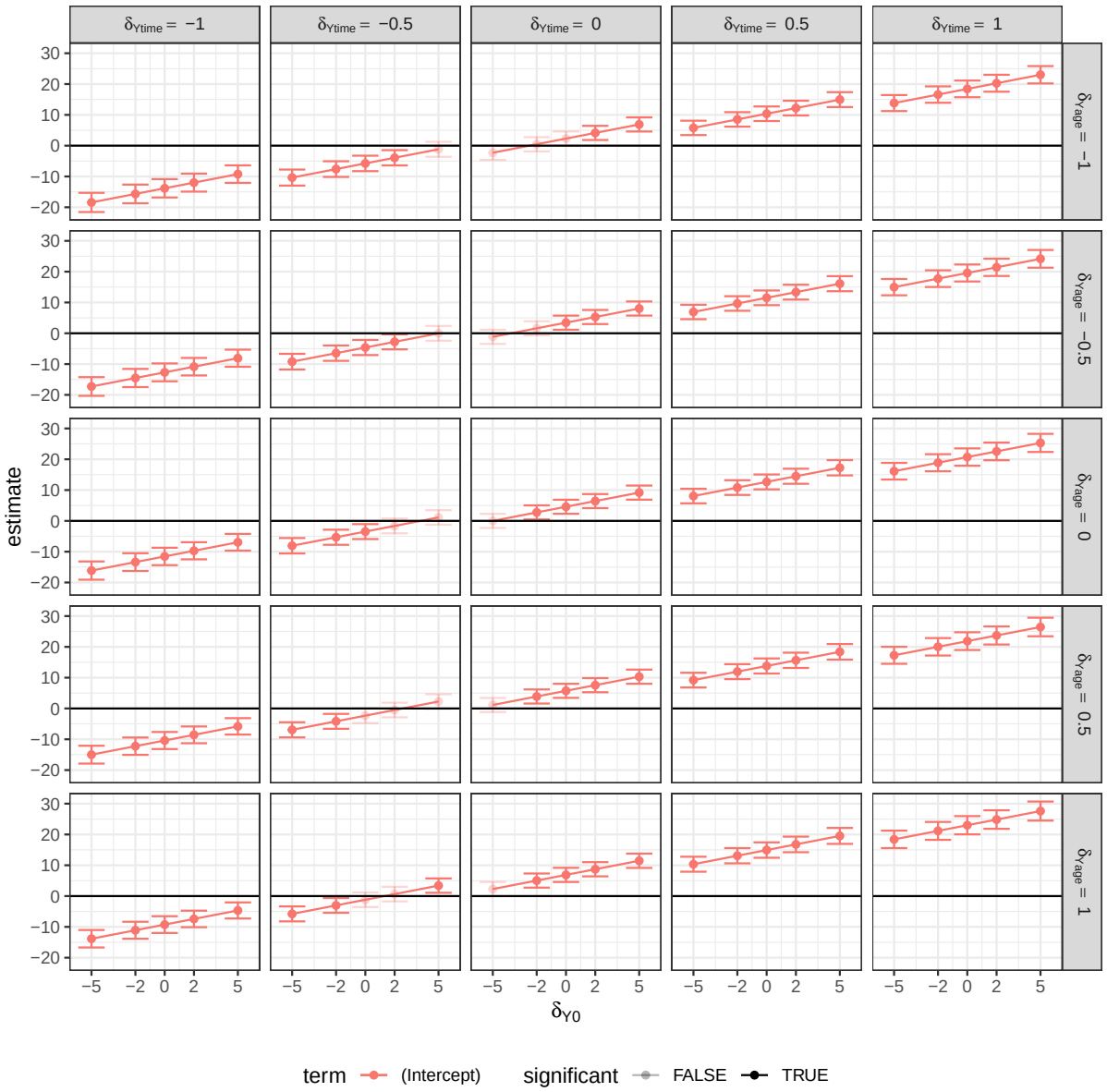

Figure 3: Sensitivity analysis of the effect on the estimated intercept and the 95 percent confidence intervals. The rows show different values for deltage. The columns show different values for deltatime. The confidence intervals are less bright if the confidence interval entails zero.

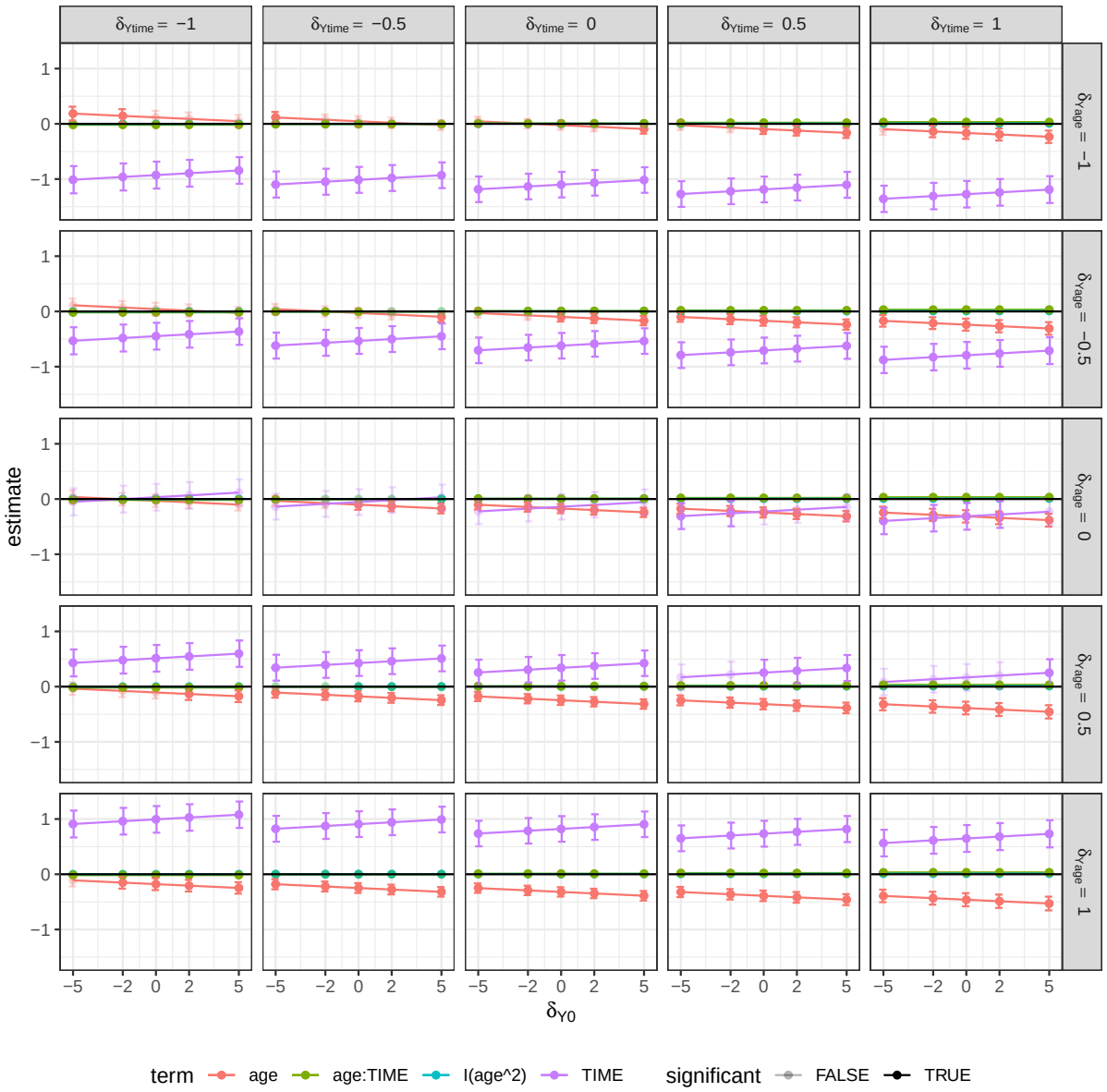

Figure 4: Sensitivity analysis of the effect on all parameter estimates, except for the intercept. The rows show different values for deltage. The columns show different values for deltatime. The 95 percent confidence intervals are less bright if the confidence interval entails zero.

of the data is imputed. It is, however, more interesting to look at the other parameter estimates. Figure 5 shows the effects on the different variables even clearer.

The parameter estimate for age is usually negative, except when δ_{Ytime} and δ_{Yage} are very negative. The interaction effect between time and age is always positive when $\delta_{Y0} >= 0$ and negative or insignificant if $\delta_{Y0} < 0$. The parameter estimate for age^2 is usually positive, unless all three δ's become negative. The parameter does not seem very impacted by δ_{Ytime} or δ_{Y0} but it is very sensitive to changes in δ_{Yage}.

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

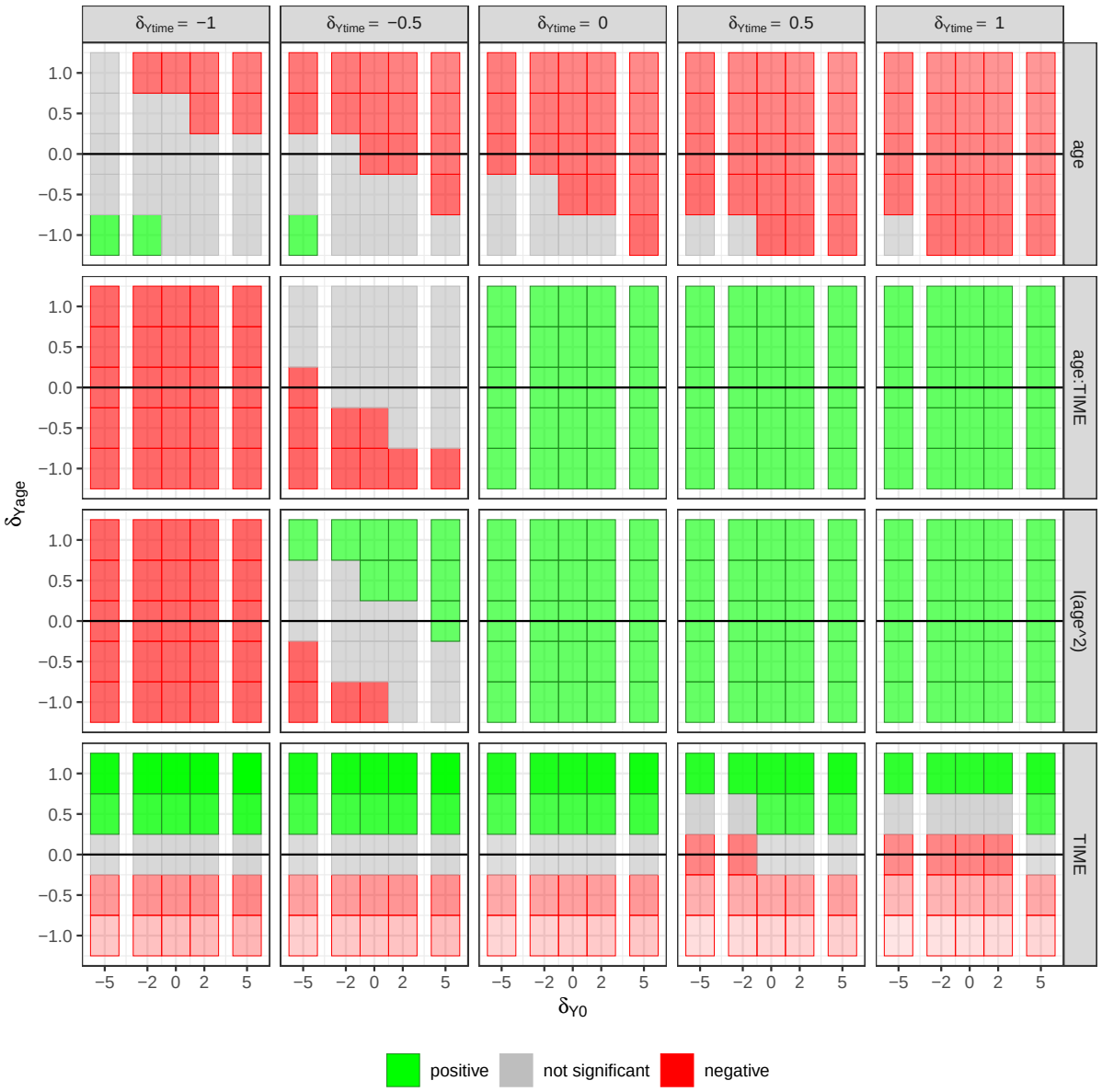

Figure 5: Heatmap for all parameter estimates, except for the intercept. The rows show the different parameters (covariates). The columns show different values for deltatime.

4 Conclusion

The hearing threshold dataset it very unbalanced and a lot of data is missing. Therefore, one needs to be very careful when interpreting models that are fit on the incomplete data.

Both direct likelihood methods and general estimating equations with and without multiple imputation are tested on the data. Generally, all models still confirm that subjects loose hearing as time passes and as the get older, regardless of the imputation method. There is also still some support for a learning effect meaning subjects seem to have worse hearing the first time they undergo the test, as was also observed in previous literature.

Any violation of MAR may have a significant impact on the model estimates, as shown in the sensitivity analysis.

5 Future research

It might be useful to replicated the current analysis but with a coarser time frame. In this report, time was discretized by rounding down the time for each measurement. That resulted in 23 time instances (year 0 until year 22) and a lot of missing data since many subjects return, on average, around every two years (81.8% missingness Figure 1). Changing the time scale to $\{0, 2, 4, ..., 22\}$ yields only 12 time instances and decreases missingness to 67.11% and the missingness comes closer to monotone missingness.

The sensitivity analysis using NARFCS may also be improved upon with the input of practitioners. Additionally, many other MNAR sensitivity analyses exist and may yield different insights.

Bibliography

Beunckens, Caroline, Geert Molenberghs, and Michael G Kenward. 2005. "Direct Likelihood Analysis Versus Simple Forms of Imputation for Missing Data in Randomized Clinical Trials." *Clinical Trials (London, England)* 2 (5): 379–86.

Brant, Larry J., and James L. Fozard. 1990. "Age Changes in Pure-tone Hearing Thresholds in a Longitudinal Study of Normal Human Aging." *The Journal of the Acoustical Society of America* 88 (2): 813–20. https://doi.org/10.1121/1.399731.

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Li, Meijuan, Nan Chen, Yang Cui, and Hongyun Liu. 2017. "Comparison of Different LGM-Based Methods with MAR and MNAR Dropout Data." *Frontiers in Psychology* 8. https://doi.org/10.3389/fpsyg.2017.00722.

Tompsett, Daniel Mark, Finbarr Leacy, Margarita Moreno-Betancur, Jon Heron, and Ian R White. 2018. "On the Use of the Not-at-Random Fully Conditional Specification (NARFCS) Procedure in Practice." *Statistics in Medicine* 37 (15): 2338–53.

Van Steen, Kristel, Geert Molenberghs, Geert Verbeke, and Herbert Thijs. 2001. “A Local Influence Approach to Sensitivity Analysis of Incomplete Longitudinal Ordinal Data.” *Statistical Modelling* 1 (2): 125–42.

Verbeke, Geert, Geert Molenberghs, Herbert Thijs, Emmanuel Lesaffre, and Michael G Kenward. 2001. “Sensitivity Analysis for Nonrandom Dropout: A Local Influence Approach.” *Biometrics* 57 (1): 7–14.