Multivariate statistics: Assignment 1

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

1 Data trichotomization

To trichotomize the data, suitable cut-off points need to be found. The cutoff points are often chosen based on either expert knowledge or so as to optimize predictive power. An easy, often used method for dichotomization is a median-split since it assures that there are an equal amount of observation at either side of the cut-off value. Similarly, for trichotomization, we could aim for approximately 33.33% of the observations in each of the three categories. That would result in the following three categories: [-12,4], (4,11], (11,70].

It is quite common in literature to dichotomize hearing loss into normal hearing (≤25 dB) and hearing loss (>25 dB) (see Garinis et al. 2017; Gallagher et al. 2019; Ju et al. 2022, for example). However, thichotomization is less common and it should be noted that it is generally not advised to discretize continuous data since some information is inevitably lost (Nelson et al. 2017; MacCallum et al. 2002).

The Centers for Disease Control and Prevention distinguishes the following levels of hearing loss, based on Clark (1981):

- ≤25 dB: Normal hearing
- 26 - 40 dB: Mild hearing loss
- 41 - 55 dB: Moderate hearing loss
- 56 - 70 dB: Moderate / severe hearing loss
- 71 - 90 dB: Severe hearing loss
- ≥91 dB: Profound hearing loss

Table 1: Number of observations in each pre-defined categories from Clark (1981).

Category	Nb observations	Percentage	Cumulative percentage	Nb subjects	Avg age
(-13,25]	4148	93.87	93.87	536	56.12
(25,40]	239	5.41	99.28	91	71.85
(40,55]	22	0.50	99.77	14	75.70
(56,70]	10	0.23	100.00	1	70.18

Table 1 shows that, in this dataset, there is no one in the severe hearing loss categories and the large majority has normal hearing (93.87%). The median for all observation with normal hearing (≤25dB) is 6 dB. We therefor suggest to trichotomize the data into the following categories:

- ≤6 dB: Excellent hearing

- 7 - 25 dB: Normal hearing
- ≥25 dB: Hearing loss

Table 2: Number of observations in each category.

Category	Nb observations	Percentage	Cumulative percentage	Nb subjects	Avg age
Excellent	2192	49.60	49.60	400	50.10
Normal	1956	44.26	93.87	414	62.88
Hearing loss	271	6.13	100.00	93	72.10

2 Methodology

As discussed in the previous section, the dependent variable will be split up into three categories. As such, the dependent variable is tranformed from a continuous (integer) variable into an ordinal one where excellent hearing is the lowest level and hearing loss is the highest.

In congruence with previous analysis, the ordinal model will be proportional odds logistic regression:

$$\begin{cases} logit[P(Y_i \leq Excellent|x_i)] = \alpha_1 + \beta_1 age_i + \beta_2 TIME_i + \beta_3 learning_i + \\ \qquad\qquad \beta_4 age_i^2 + \beta_5 age_i * TIME_i \\ logit[P(Y_i \leq Normal|x_i)] = \alpha_2 + \beta_1 age_i + \beta_2 TIME_i + \beta_3 learning_i + \\ \qquad\qquad \beta_4 age_i^2 + \beta_5 age_i * TIME_i \end{cases} \tag{1}$$

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

3 Results

3.1 Marginal model

First, we fit a marginal model with the *ordLORgee* from the **multgee**. This function allows for an ordinal dependent variable which is appropriate for our data. The result is shown in Table 3.

3.2 Random-effects model

On top of the fixed effects (equation (1)), the random effects model only includes a random intercept since it did not converge with random slopes included. The covariate age was also standardized and centered to improve convergence.

The random intercept has a variance (standard deviation) of 5.68 (2.38).

Table 3: Estimated GEE model

Parameter	Estimate	Odds
α_1	0.68	1.9790957
α_2	3.87 ***	47.8628677
age	0.04 .	1.0445853
TIME	0.01	1.0149608
learning	-0.24 **	0.7849463
age^2	-0.00 ***	0.9991204
age:TIME	-0.00 ***	0.9981717

Table 4: Estimated mixed effects model

Parameter	Estimate	Odds
α_1	0.97 ***	2.642926
α_2	6.45 ***	633.126412
age$_{scale}$	1.44 ***	4.219361
TIME	0.13 ***	1.142768
learning	0.39 **	1.480481
age$^2_{scale}$	0.36 **	1.437232
age$_{scale}$*TIME	0.05 ***	1.050161

3.2.1 Empirical Bayes prediction

To infer the marginal evolution of hearing loss over time, one cannot simply set the random intercept equal to zero to calculate the prediction. This is because the expectation of a logit function is not equal to the logit of the expectation. Figure 1 shows the evolutions for the average subjects of a certain age (where $b_i = 0$) to the marginal evolutions (integrated GLMM). The latter are used to get marginal predictions for each of the ordinal levels in Figure 2. This figure clearly shows the learning effect at $TIME == 0$. The youngest subject have the highest probability of having excellent hearing and that probability goes down as time progresses. For older subjects, there seems to be a tipping point where the probability of having normal hearing starts to go down, as the probability of having hearing loss steeply increases.

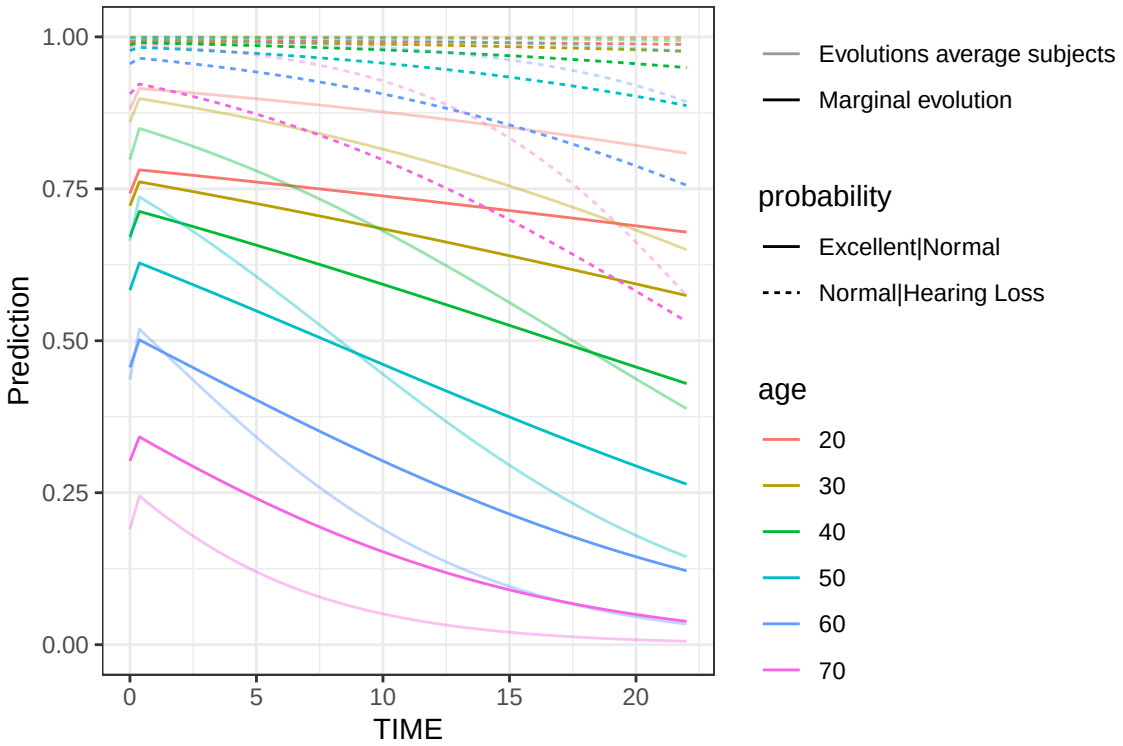

Figure 1: Marginal predictions versus the evolutions for the average subjects

Apart from the marginal evolution, we can also get empirical bayes estimates for all subjects. Figure 3 shows the distribution of the random intercept on the left and the scatterplot on the right shows how the random intercepts related to the subject's age and follow-up time. It can be seen that there are more outliers on the right, meaning that these subjects have higher than expected hearing threshold i.e. a higher than expected probability of hearing loss. There are 7 subjects with EB estimate > 5. 5 of these subjects were followed up more than 10 years and they range from age 39 to 66.2. In the current dataset, no clear reason can be find as to why these subjects deviate so much from expectation. It would be interesting to try to link the EB estimates to other characteristics that might influence hearing such as occupation for example.

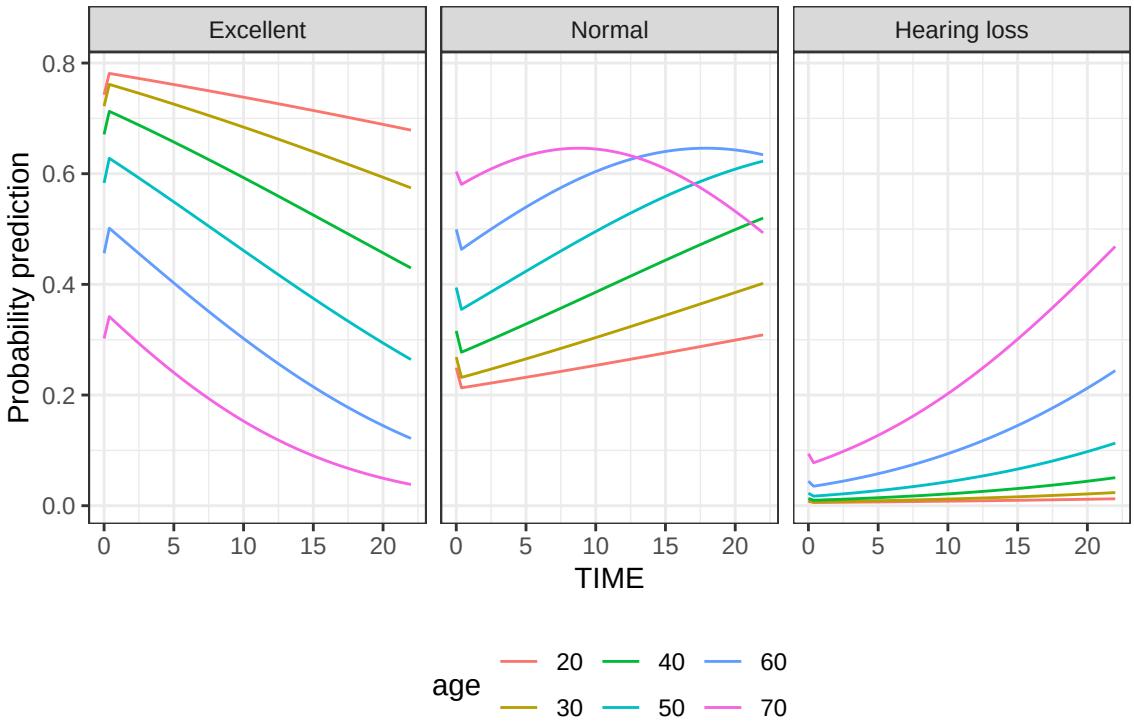

Figure 2: Marginal evolutions

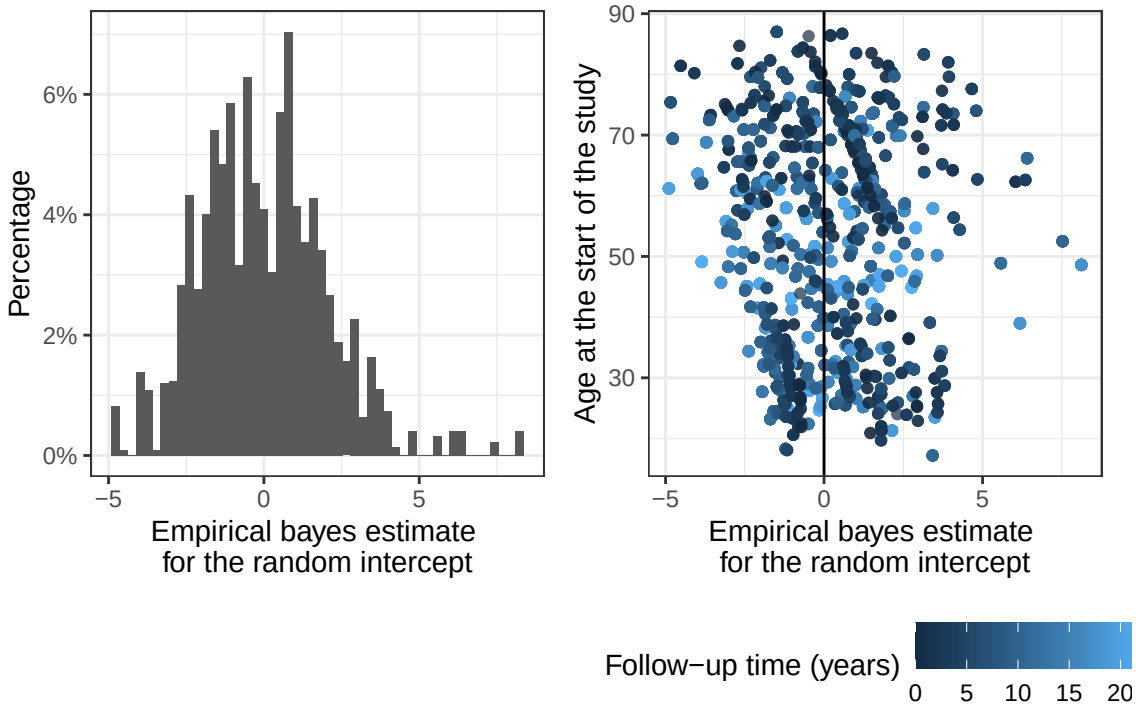

Figure 3: Empirical Bayes (EB) estimates

3.3 Transition model

4 Discussion

Bibliography

Clark, JG. 1981. “Uses and Abuses of Hearing Loss Classification.” *ASHA* 23 (7): 493–500.

Gallagher, Nicola E., Chris C. Patterson, Charlotte E. Neville, John Yarnell, Yoav Ben-Shlomo, Anne Fehily, John E. Gallacher, Natalie Lyner, and Jayne V. Woodside. 2019. “Dietary Patterns and Hearing Loss in Older Men Enrolled in the Caerphilly Study.” *British Journal of Nutrition* 121 (8): 877–86. https://doi.org/10.1017/S0007114519000175.

Garinis, Angela C, Campbell P Cross, Priya Srikanth, Kelly Carroll, M Patrick Feeney, Douglas H Keefe, Lisa L Hunter, et al. 2017. “The Cumulative Effects of Intravenous Antibiotic Treatments on Hearing in Patients with Cystic Fibrosis.” *Journal of Cystic Fibrosis* 16 (3): 401–9.

Ju, Min Jae, Sung Kyun Park, Sun-Young Kim, and Yoon-Hyeong Choi. 2022. “Long-Term Exposure to Ambient Air Pollutants and Hearing Loss in Korean Adults.” *Science of The Total Environment* 820: 153124.

MacCallum, Robert C, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. 2002. “On the Practice of Dichotomization of Quantitative Variables.” *Psychological Methods* 7 (1): 19.

Nelson, S. L. Prince, V. Ramakrishnan, P. J. Nietert, D. L. Kamen, P. S. Ramos, and B. J. Wolf. 2017. “An Evaluation of Common Methods for Dichotomization of Continuous Variables to Discriminate Disease Status.” *Communications in Statistics - Theory and Methods* 46 (21): 10823–34. https://doi.org/10.1080/03610926.2016.1248783.