FACULTY OF BIOSCIENCE ENGINEERING SCIENCE
BIOINFORMATICS
KASTEELPARK ARENBERG 20 3001 LEUVEN

Statistical Methods for Bioinformatics

Report

Case Study:

Group 1: Rita Andrade (r0927665)
Zofia Urbaniak (r0927663)
Hanne Pelemans (r0808530)
Tarek Chammaa El Rifai Mashmoushi (r0977298)
Disha Sureesh(r0964352)

August 2024

1 Introduction

1.1 Study design-Objective

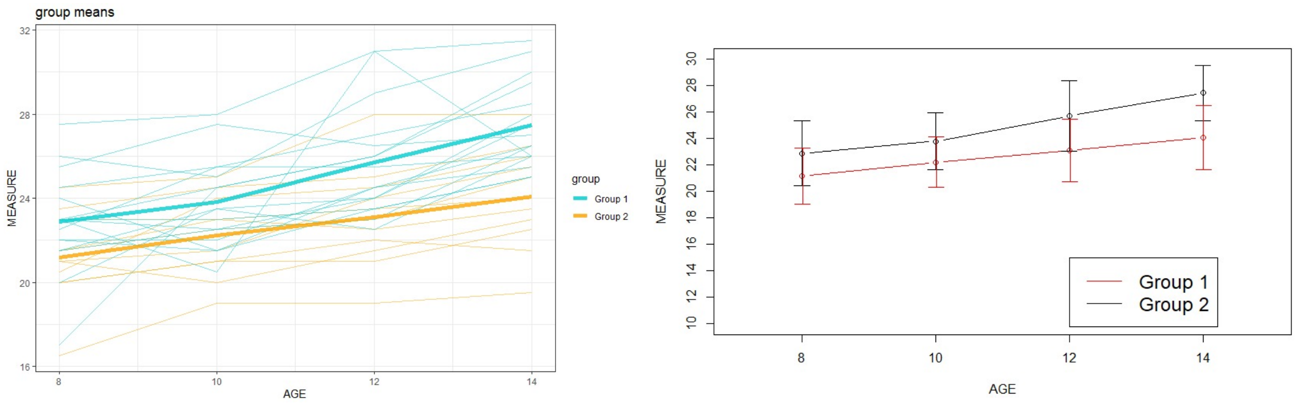
This report tackles the scientific question whether dental growth is related to gender or not. The data used is from the study in Biometrika (1964) reported by Potthoff and Roy. This study recorded the distance from the center of the pituitary to the maxillary fissure for 16 boys (group 1) and 11 girls (group 2). The measurements were taken at ages 8, 10, 12 and 14. So the response variable is the distance from the center of the pituitary to the maxillary fissure and the two different variables are sex and age. In what follows, the results of a longitudinal study on these data are discussed. First, the descriptive statistics are discussed. Then a model is constructed and evaluated, taking into account the fixed effects, random effects and residual error. Finally, that model is used to assess the data and draw conclusions. The tables and graphs seen in this report come from the attached script.

1.2 Descriptive statistics

To begin with, the data is reviewed and checked for missing data. In this study, there is none. On overview of the data was made (Table 1) with the number of observations, the mean and the standard deviation per group. Then the data was visualised in a spaghetti plot (Figure 1). This shows the distance in function of the age for the boys (group 1) in blue and for the girls (group 2) in orange. The averages for the boys and girls are the bold lines. The starting points for the distance of the two groups are mostly in the same range. This is because the subjects per group are randomized. In the plot, it can be seen that for both sexes, the distance increases with age. It makes sense that the distance becomes greater when they grow older. It can thus be expected that the distance is depends on the age. In addition, it is noticeable that from the age of 10, the distance for the boys increases faster than before while for the girls it continues to grow seemingly linearly. From this it can be deduced that the distance probably also depends on the variable sex. The box plots in the script show the same trends. The mean evolution of the distance from the center of the pituitary to the maxillary fissure with regard to sex and age is positive.

AGE		8			10			12			14		
Statistics		n	Mean	sd	n	Mean	sd	n	Mean	sd	n	Mean	sd
SEX	1	16	22.87500	2.452889	16	23.81250	2.136001	16	25.71875	2.651847	16	27.46875	2.085416
	2	11	21.18182	2.124532	11	22.22727	1.902152	11	23.09091	2.364510	11	24.09091	2.437398

Table 1: Descriptive Statistics



(a) Spaghetti plot with mean evolution of all participants

(b) Mean evolution per group

Figure 1: Mean evolution

1.3 Correlation matrix

As final descriptive statistic, the correlation matrix was calculated (Table 2). The matrix shows that the response values were more strongly correlated for the years that are closer together. The measurements taken further apart however, are less related to each other. Both trends are as expected.

	MEASURE.8	MEASURE.10	MEASURE.12	MEASURE.14
MEASURE.8	1.0000000	0.6255833	0.7108079	0.5998338
MEASURE.10	0.6255833	1.0000000	0.6348775	0.7593268
MEASURE.12	0.7108079	0.6348775	1.0000000	0.7949980
MEASURE.14	0.5998338	0.7593268	0.7949980	1.0000000

Table 2: Correlation between mean measurements at each age

2 Mixed Effects Model

The goal is to model the evolution over time for specific individuals as well as to model the differences between the subjects in order to draw conclusions about the impact of gender on dental growth .

2.1 Level 1

In the first level, we will perform linear regression on each participant to model the individual time evolution with a linear function. We will allow for different intercepts and slopes for each individual. The level 1 model is:

$$Y_{ij} = \pi_{0i} + \pi_{1i} \cdot AGE_{ij} + e_{ij} \tag{1}$$

The histogram of individual intercepts suggests that the values are normally distributed, supporting the effective randomization of our study. The histogram of individual slopes indicates that the changes over time are predominantly positive, which aligns with the expected consistent dental growth with age. Lastly, the histogram for R-squared values reveals that for most participants, the model provided a good fit, explaining a significant portion of the variation (Figure 2). Thus, for the purposes of this analysis, there is no need to go beyond a linear model.

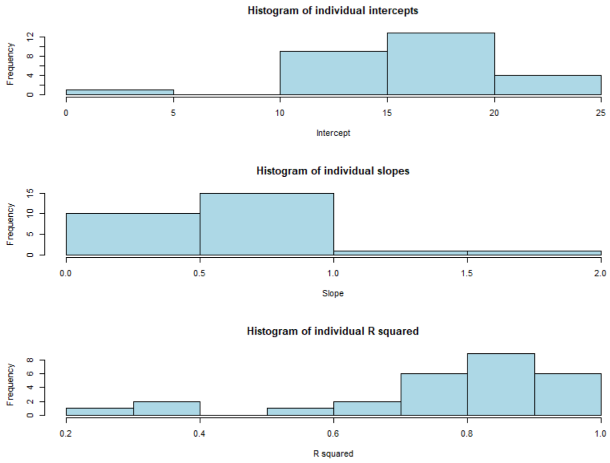


Figure 2: Histograms

2.2 Level 2

Now we will attempt to explain the variability in the intercepts and evolution between the individuals. For this purpose, we will utilize the following level 2 model with γ_{00} being the average intercept and γ_{10} the average slope. Additionally the b_{0i} and b_{1i} are the error terms.

$$\pi_{0i} = \gamma_{00} + \gamma_{01} \cdot \mathrm{SEX}_i + b_{0i} \tag{2}$$

$$\pi_{1i} = \gamma_{10} + \gamma_{11} \cdot \mathrm{SEX}_i + b_{1i} \tag{3}$$

After, we consider the following covariance matrix, adding the σ_{01} to account for the possibility that the intercept π_{01} and the slope π_{1i} are correlated.

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right) \tag{4}$$

3 Model Estimation

The main model is set as a mixed effect model fitted with the lmer function and Maximum Likelihood for parameter calculation. We compared 4 models and will now interpret the results.

3.1 Fixed Effects

We ran 4 models: one that doesn't account for the effect of SEX ($\gamma_{01} = \gamma_{11} = 0$), one with the interaction between AGE and SEX (the full model), one with the interaction and AGE, and one with both variables and the interaction. Always with the response variable "MEASURE". We compared the models using ANOVA and found that the second model is the best one, as adding both the main effect of SEX and AGE doesn't add an improvement to our full model. Looking at our chosen model and the summary results:

- The main affect is AGE (p-value = 4.42e-10). There's a significant positive correlation between the response variable and the age of the child.
- The interaction of sex with age has a smaller significant impact on the growth registered (p-value = 0.0263). The effect of age on the measure does not vary a lot between the two sex groups.
- The sex of the sample doesn't seem to have an influence on the response variable (p-value = 0.5072).

The null hypothesis is that there is no difference between the effects of both sexes on MEASURE over time. This hypothesis is rejected when we look into the ANOVA results. We can see that adding the fixed effects of the SEX variable doesn't produce a significant change to the full model. With these results the level 2 of our model looks like:

$$\pi_{0i} = 16.34062 + b_{0i} \tag{5}$$

$$\pi_{1i} = 0.78438 + b_{1i} \tag{6}$$

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: MEASURE ~ 1 + SEX * AGE + (1 + AGE | IDNR)
   Data: data

     AIC      BIC   logLik deviance df.resid
   443.8    465.3   -213.9    427.8      100

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3360 -0.4154  0.0104  0.4917  3.8582

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 IDNR     (Intercept) 4.55683  2.1347
          AGE         0.02376  0.1541   -0.60
 Residual             1.71621  1.3100
Number of obs: 108, groups:  IDNR, 27

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) 16.34062    0.98008 26.99974  16.673 9.70e-16 ***
```

```
SEX2          1.03210    1.53549 26.99974   0.672   0.5072
AGE           0.78438    0.08275 26.99979   9.479 4.42e-10 ***
SEX2:AGE     -0.30483    0.12965 26.99979  -2.351   0.0263 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr) SEX2   AGE
SEX2     -0.638
AGE      -0.880  0.562
SEX2:AGE  0.562 -0.880 -0.638
```

Models	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
data.lmer2.nosex	6	451.21	467.30	-219.61	439.21			
data.lmer2.intsex	7	446.84	465.61	-216.42	432.84	6.3764	1	0.01156 *
data.lmer2	8	443.81	465.61	-213.90	427.81	5.0292	1	0.02492 *

Table 3: Anova

3.2 Random Effects

In examining the random effects, we note a negative correlation (-0.60) between the random intercept and the random slope. This implies that children in the study who initially exhibit higher measurements of distance from the center of the pituitary to the maxillary fissure (i.e., higher intercept) tend to undergo slower rates of growth (gentler slopes) compared to participants with initially smaller measurements of the same parameter.

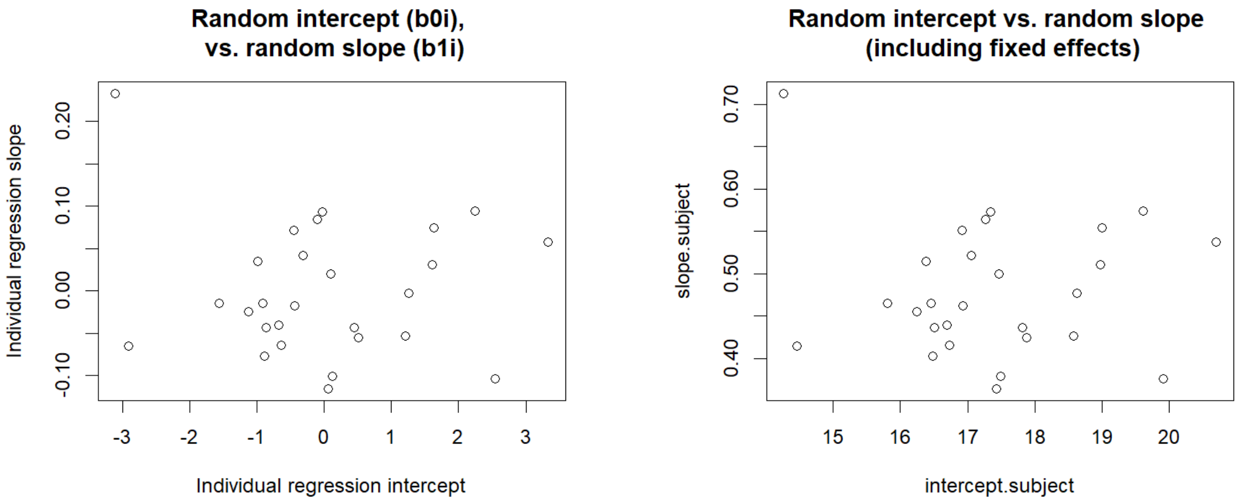


Figure 3: Correlation graphs for random intercept and slope