Generalised Regression Models

GRM: Case Study — GLMMs **Semester 1, 2022–2023**

FEEDBACK

Initial R analysis

Download and load the data into R:

```
dental <- read.csv(file = "dental.csv")
```

Fit a simple linear regression model for distance with only the age covariate:

```
model_initial <- lm(formula = distance ~ age, data = dental)
summary(model_initial)

Call:
lm(formula = distance ~ age, data = dental)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5037 -1.5778 -0.1833  1.3519  6.3167

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.7611     1.2256  13.676  < 2e-16 ***
age           0.6602     0.1092   6.047 2.25e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.537 on 106 degrees of freedom
Multiple R-squared:  0.2565,    Adjusted R-squared:  0.2495
F-statistic: 36.56 on 1 and 106 DF,  p-value: 2.248e-08
```

From the estimates, we see that the average growth rate is estimated to be 660µm per year.

Repeating the investigation but for only child with `"ID08"` gives:

```
dental_ID08 <- dental[dental[,"id"] == "ID08", ]
model_ID08 <- lm(formula = distance ~ age, data = dental_ID08)
summary(model_ID08)

Call:
lm(formula = distance ~ age, data = dental_ID08)

Residuals:
   29    30    31    32
 0.15 -0.20 -0.05  0.10

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.4500     0.4861  44.131 0.000513 ***
age           0.1750     0.0433   4.041 0.056120 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

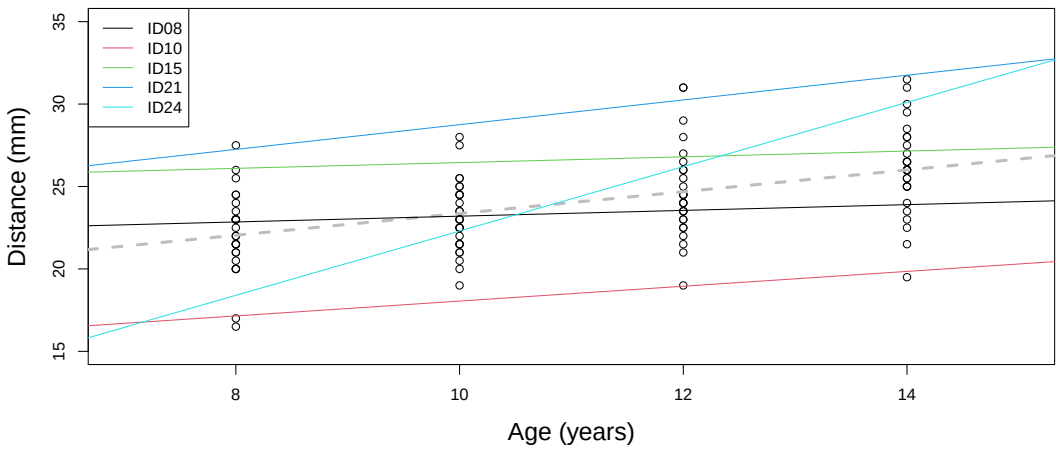
```
Residual standard error: 0.1936 on 2 degrees of freedom
Multiple R-squared:  0.8909,    Adjusted R-squared:  0.8364
F-statistic: 16.33 on 1 and 2 DF,  p-value: 0.05612
```

The growth rate for this child is slower than the average across the study.

Quiz 1 & Quiz 2: Repeating the above analysis for children with IDs `"ID10"`, `"ID15"`, `"ID21"` and `"ID24"` can be performed concisely as follows.

```
plot(dental[, "age"], dental[,"distance"])
abline(a = 16.7611, b = 0.6602, col = "grey", lwd = 3, lty = 2)
Selected_IDs <- c("ID08","ID10","ID15","ID21","ID24")
Estimates <- matrix(0, nrow = 2, ncol = length(Selected_IDs))
for(i in seq_along(Selected_IDs)){
  dental_ID <- dental[dental[,"id"] == Selected_IDs[i], ]
  model_ID  <- lm(formula = distance ~ age, data = dental_ID)
  Estimates[,i] <- coef(model_ID)
  abline(a = Estimates[1,i], b = Estimates[2,i], col = i)
}
legend("topleft", legend = Selected_IDs, col = 1:5, lty=1)
colnames(Estimates) <- Selected_IDs
rownames(Estimates) <- c("(Intercept)","age")
Estimates
              ID08  ID10   ID15  ID21 ID24
(Intercept) 21.450 13.55 24.700 21.25 2.80
age          0.175  0.45  0.175  0.75 1.95
```

All of the growth rate estimates are positive, but they are notably different for each child. This is clear from the plot with some children having a shallower gradients and others with stepper gradients than the overall average (grey dashed line). `model_inital` is not a good model as there is more variation in the data than what the model is able to describe.



Fitting a linear mixed model

Load the package `lme4` that contains the functions that we need to fit a linear mixed effects model:

```
library(lme4)
```

(Note: If you don't have this package then you must first run `install.packages("lme4")`.)

Fit the linear mixed model with random slope on each child's age:

```
model_lmm <- lmer(formula = distance ~ age + (-1 + age | id), data = dental)
summary(model_lmm)

Linear mixed model fit by REML ['lmerMod']
Formula: distance ~ age + (-1 + age | id)
   Data: dental
REML criterion at convergence: 445.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9317 -0.4521  0.0026  0.4834  3.6399

Random effects:
 Groups   Name Variance Std.Dev.
 id       age  0.03593  0.1895
 Residual      1.99555  1.4126
Number of obs: 108, groups:  id, 27

Fixed effects:
            Estimate Std. Error t value
(Intercept)  16.7611     0.6824  24.563
age           0.6602     0.0709   9.312

Correlation of Fixed Effects:
    (Intr)
age -0.840
```

Quiz 3: From `model_initial` the estimated standard error the age coefficient is $s.e.(\hat{\beta}_1) = 0.1092$, and the corresponding estimate from `model_lmm` is $s.e.(\hat{\beta}_1) = 0.0709$. This means that the standard error in $\hat{\beta}_1$ is lower for the linear mixed model.

Quiz 4: From the summary of `model_inital` we see that the residual standard error is 2.537. From the random effects table of the `model_lmm` summary we see that the residual standard error is lower at 1.4126.

Comparing models

Compare the simple linear model and the linear mixed model:

```
anova(model_inital, model_lmm)

refitting model(s) with ML (instead of REML)
Data: dental
Models:
model_initial: distance ~ age
model_lmm: distance ~ age + (-1 + age | id)
              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model_initial    3 511.58 519.62 -252.79   505.58
model_lmm        4 449.45 460.18 -220.73   441.45 64.123  1  1.169e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The linear mixed model has a lower AIC, and the p-value of the chi-squared deviance test is a very small. Therefore the linear mixed model is better in describing the variation in dental distances.

Quiz 5: Fit a linear mixed model with both random intercept and random age slope for each child in the study:

```
model_lmm2 <- lmer(formula = distance ~ age + (age | id), data = dental)
anova(model_lmm2, model_lmm)
```

```
refitting model(s) with ML (instead of REML)
Data: dental
Models:
model_lmm: distance ~ age + (-1 + age | id)
model_lmm2: distance ~ age + (age | id)
           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model_lmm     4 449.45 460.18 -220.73   441.45
model_lmm2    6 451.21 467.30 -219.61   439.21 2.2427  2     0.3258
```

From the ANOVA investigation, we can see that the two models are statistically equivalent according to the chi-squared deviance test as the p-value of 0.3258 is larger than the 5% significance level. Therefore we prefer the simpler description given by the random slope only model. This is also reflected with a lower AIC for `model_lmm`.