University of Edinburgh
School of Mathematics

Generalised Regression Models

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

Problem and Data

In this case study we shall be examining the data in `dental.csv` which is from Potthoff & Roy[1] who investigated dental growth rate amongst children. The study consists of 27 children who were monitored from the age of 8 to 14. At each assessment, the distance between the pituitary gland to the pterygomaxillary fissure was measured. The file contains the following variables:

Variable	**Description**
id	A unique identifier for the 27 children in the study.
gender	Gender of the child identified at birth (`M/F`).
age	Child's age at which each of the 4 measurements were taken (`8, 10, 12 & 14`).
distance	The measured dental distance (in mm).

Our goal is to construct a regression model to best evaluate the dental growth rate, whilst also accounting for individual variation between the children.

Initial R analysis

Download and save the data file and read it into R:

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

To begin, create a simple linear regression model for distance with age as the only covariate:

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

From the estimates, the linear predictor is $\mathbb{E}[Y] = 16.76 + 0.660x$, meaning that the distance between the pituitary gland to the pteryomaxillary fissure increases on average by 660μm per year between the ages of 8 and 14. However, is this growth rate the same for all children in the study? Extract from `dental` the data for the child with `"ID08"` and evaluate the simple linear model for this child:

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

The linear predictor for this child is $\mathbb{E}[Y] = 21.45 + 0.175x$, so her growth rate of 175μm per year is slower than the average across the children in the study.

Quiz 1: Repeat the above investigation for the children with IDs `"ID10"`, `"ID15"`, `"ID21"` and `"ID24"`.

Quiz 2: Create a scatter plot of distance against age and add the linear predictor from `model_initial` and for the five children (Hint: `help(abline)`). Describe what you see. Do you think `model_inital` is a good model?

Statistical Model

We wish to fit a linear mixed model to account for the variation in growth rate amongst the individual children within the study. Formally, let Y_{ij} be the dental distance for the i^{th} child ($i = 1, \ldots, 27$) at their j^{th} assessment (for $j = 1, \ldots, 4$), and denote the child's age at this time by x_{ij}. The linear mixed model that we are going to examine contains a random slope effect as follows:

$$Y_{ij} = \beta_0 + (\beta_1 + u_i)x_{ij} + \epsilon_{ij}. \tag{1}$$

[1] Potthoff, R. F., Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51(3), 313-326.

The parameters β_0 & β_1 are the intercept and slope parameters that is common for all children, and so are called the fixed effects. The u_i's are the random slope component (also known as a random effect) that adjusts the slope of the regression model for each child. It is assumed that $u_i \sim N(0, \sigma_u^2)$ for $i = 1, \dots, 27$, where σ_u^2 describes the variance, or spread, amongst the random effects. Finally, we assume that all of the error terms are normal, and satisfy $\mathbb{E}[\epsilon_{ij}] = 0$ and $\text{Var}(\epsilon_{ij}) = \sigma_\epsilon^2$.

Fitting a linear mixed model

The functions we need to fit a linear mixed model are in the `lme4` package[2], which we can load with:

```
library(lme4)
```

A linear mixed model is fitted using the `lmer()` command. The inputs for this function are identical to `lm()`, namely:

`formula`	A symbolic description of the model to be fitted.
`data`	A data frame containing the variables in the model.

However, the syntax for `formula` is slightly different:

```
    <Response variable> ~ <Fixed effects> + (<Random effects> | <grouping>)
```

The first half is identical to the `y~x` structure used for `lm()` and represents the fixed effect components for the linear mixed model. The additional bracketed component encodes the random effects for the model. The terms before the vertical bar, `|`, are the covariates for the random effects (e.g. child's age, `age`) and the term after the vertical bar states which variable defines the grouping (e.g. child's ID, `id`).

Copy and run the following code that fits the linear mixed model specified by equation (1):

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

Note: An intercept term for both fixed and random components are specified by default (see `help(formula)`). The inclusion of `-1` inside the bracket removes the default random intercept so that the resulting model *only* has a random slope with respect to the `age` covariate.

The summary from the fitted linear mixed model contains two tables titled `Fixed effects` and `Random effects`. The structure of the fixed effects table is similar to the output from a linear model and contains the estimates for the intercept and age covariate parameters β_0 and β_1. From comparing this table to the summary of `model_inital` we see that the estimates are identical in fact identical: $\hat{\beta}_0 = 16.76$ and $\hat{\beta}_1 = 0.660$. However, their estimated standard errors are different.

Quiz 3: Which model the lowest estimated standard error for the age fixed effect coefficient, β_1?

The table titled `Random effects` in the summary contains the estimates for the two variance parameters. From this we can see that the estimated variance for the random slope effects, i.e. amongst the u_i's in equation (1), is $\hat{\sigma}_u^2 = 0.0359$.

Quiz 4: From the summary of both `model_inital` and `model_lmm`, find the residual standard error. Which model has the lowest value?

Comparing models

From the estimated residual standard error we see that the linear mixed model with random slope on the age of the children accounts for more variation than a simple linear model. But is the additional model complexity of linear mixed model necessary or is the simple linear model statistically sufficient to describe the data? To assess this we can perform an ANOVA investigation between the two models using the `anova()` command:

```
anova(model_inital, model_lmm)
```

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

From the results we can see that the AIC for `model_lmm` is much smaller than for `model_inital`. Furthermore, the p-value of the chi-squared deviance test is very small at 1.17×10^{-15}, suggesting that it is highly unlikely that the two models are statistically equivalent. Therefore, we can conclude that the fitted linear mixed model is better than the simple linear regression model in explaining the variation seen in the data.

Quiz 5: Run the following to fit a linear mixed model with random intercept and random slope:

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

Use `anova()` to determine whether adding a random intercept into the model further improves the linear mixed model.

Extra: Other things to try!

So far, the covariate for the gender of the children identified at birth has not be used. Investigate different ways of adding this variable into you model to see if you can obtain a better model than `model_lmm`. You may consider adding `gender` as a fixed and or random effect, or perhaps as an interaction with `age`.