Homework 5 - BIOS 6643

Dominic Adducci

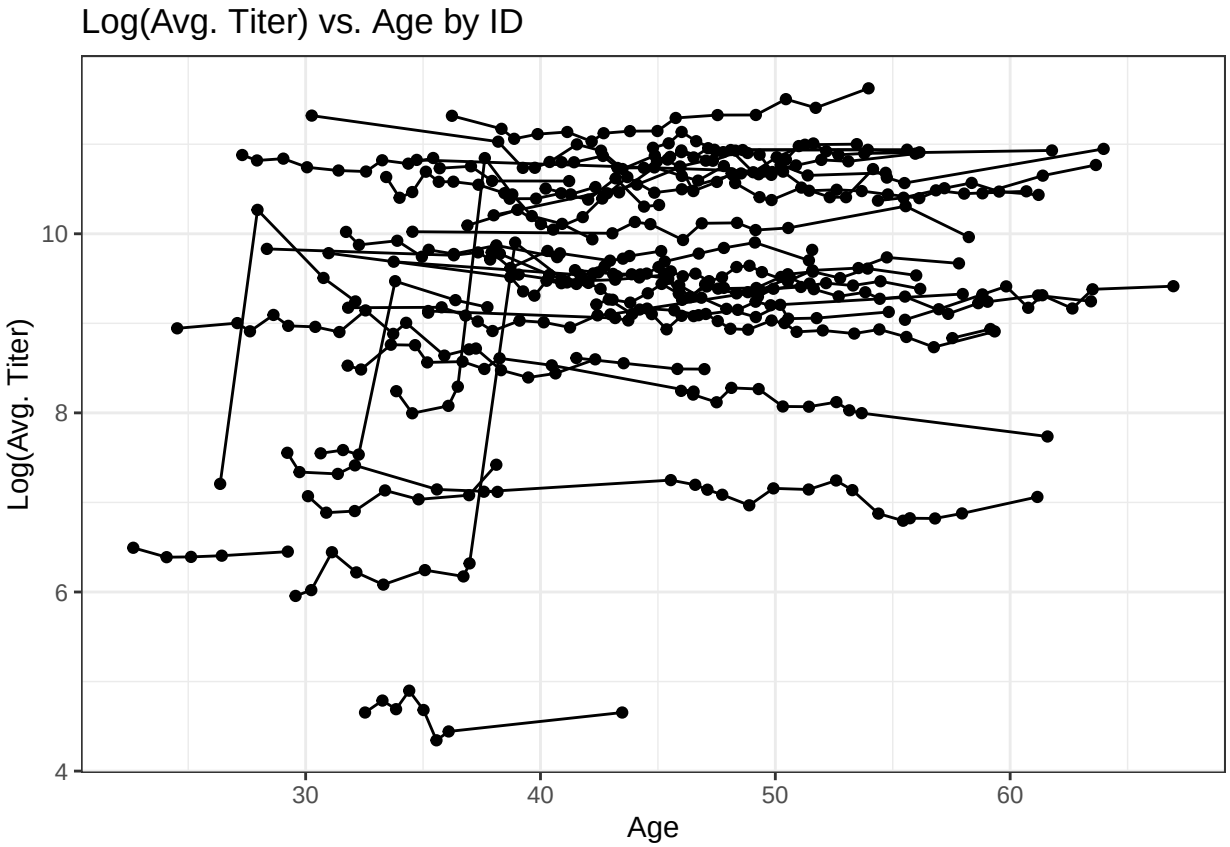
2023-09-28

We received data on the blood levels of measles vaccine titers collected randomly over an average of 12 years on 39 subjects. The science of interest was to estimate whether many decades after measles vaccination there was still a significant decay of the vaccine titers. This information will be used to determine whether boosters are needed.

Question 1

Conduct a preliminary data investigation on the titer data. Analysis will be conducted on the log scale. Please make graphs on this scale. Include in your homework submission 2-3 graphs and a summary table that you believe describes the data. Interpret both the graphs and the summary data in a paragraph.

```
## New names:
## Rows: 455 Columns: 4
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): ID dbl (3): ...1, Age.at.Draw, Avg..Titer
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```



Question 2: Random Intercept Model

Part A Write out the random intercept model for these data in subject level notation including indices for matrices.

$$Y_i = X_i\beta + Z_i b_i + E_i$$

Where each matrix has the following dimensions:

- $Y_i : (n_i \times 1)$; n_i refers to the number of time points (age). Outcome is the log of average titer.
- $X_i : (n_i \times p)$; n_i refers to the number of time points(age) and p is the number of covariates plus the intercept. The only covariates are the number of time points (age).
- $\beta : (p \times 1)$; p refers to the number of covariates (measurements at a certain age plus the intercept).
- $Z_i : (n_i \times q)$; n_i refers to the number of time points (age). q refers to the number of covariates for the random effects. In this case there is only a random intercept, so $q = 1$.
- $b_i : (q \times 1)$; q refers to the number of covariates for the random effects. In this case there is only one random intercept, so $q = 1$.
- $E_i : (n_i \times 1)$; n_i refers to the number of time points (age). Each individual outcome has its own error term.

Part B What is the format for G_i and R_i that will be assumed in the lmer R function when you fit the random intercept model? Please write the format for these two matrices and interpret.

The variability of the random effect will be a matrix with $(q_i \times q_i)$ dimensions. Because there is only one random effect, the random intercept, this will be a matrix with a single element for all subjects.

$$G_i = [\tau_0^2] = [\sigma_0^2]$$

It is usually not possible to estimate both G and an unstructured R_i, meaning that $R_i = \sigma_e^2 I_n$ (an independence matrix). The dimensions of R_i are $(n_i \times n_i)$, which in this case means a square matrix where n_i is the number of time points (measurements at different ages) for each subject. Because different subjects have different time points the dimensions of the R_i matrix will change between subjects.

Part C Will G_i and R_i be the same dimension for each individual? Explain.

For G_i the dimensions will be the same between all subjects because each subject only has a random intercept. The dimension of G_i are $(q_i \times q_i)$, where q_i refers to the number of random effects. Because this model only includes a random intercept q_i will be equal to 1, and the matrix will simply contain a single element of $\tau_0^2 = \sigma_0^2$.

R_i will not be the same between subjects because there are different numbers of time points for each subject. The dimensions of R_i are $(n_i \times n_i)$, where n_i refers to the number of individual time points (age at sample collection).

Part D Fit a random intercept model for these data using the lmer R function with REML and ML estimation approaches. Provide you code and output.

```
### START QUESTION 2 CODE ###

## START QUESTION 2 Part D CODE ##

# Fitting a lmer model with a random intercept using the REML method
measles_ri_reml <- lmer(Avg_Titer_Log ~ Age_At_Draw + (1|ID),
                        data = measles_raw, REML = TRUE)

summary(measles_ri_reml)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Avg_Titer_Log ~ Age_At_Draw + (1 | ID)
##    Data: measles_raw
##
## REML criterion at convergence: 524.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.4095 -0.2623 -0.0105  0.2590  8.7896
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 2.0834   1.4434
##  Residual             0.1154   0.3396
## Number of obs: 454, groups:  ID, 40
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 9.098e+00  2.735e-01 7.657e+01  33.267   <2e-16 ***
```

```
## Age_At_Draw 4.924e-03  3.388e-03 4.253e+02   1.453    0.147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## Age_At_Draw -0.547
```

```
# Fitting a lmer model with a random intercept using the ML method
measles_ri_ml <- lmer(Avg_Titer_Log ~ Age_At_Draw + (1|ID),
                      data = measles_raw, REML = FALSE)

summary(measles_ri_ml)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: Avg_Titer_Log ~ Age_At_Draw + (1 | ID)
##    Data: measles_raw
##
##      AIC      BIC   logLik deviance df.resid
##    521.5    538.0   -256.7    513.5      450
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.4158 -0.2623 -0.0103  0.2596  8.7988
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 2.0299   1.4247
##  Residual             0.1151   0.3392
## Number of obs: 454, groups:  ID, 40
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 9.097e+00  2.709e-01 7.949e+01  33.579   <2e-16 ***
## Age_At_Draw 4.950e-03  3.384e-03 4.266e+02   1.463    0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## Age_At_Draw -0.551
```

```
## FINISH QUESTION 2 PART D CODE ##
```

Part E What are the fixed effects estimates with 95%CI and p-values for both fits? How do they compare? Explain.

Term	Estimate	95% Conf.Low	95% Conf.High	P-Value
REML				
Intercept	9.09765	8.55305	9.64226	< 2e-16
Age At Draw	0.00492	-0.00174	0.01158	0.14692
ML				
Intercept	9.09653	8.55736	9.63569	< 2e-16
Age At Draw	0.00495	-0.00170	0.01160	0.14424

The REML method has a slightly higher intercept coefficient and a slightly lower "Age At Draw" coefficient compared to the ML method. The REML method also has a slightly wider 95% CI for the intercept term and a slightly narrower 95% CI for the "Age At Draw" term compared to the ML method. The p-value for the intercept term is effectively the same and significant between both methods. The p-value for the "Age At Draw" term is slightly lower for the ML method, but neither REML or ML have a significant "Age At Draw" coefficient. Because the ML method tends to have a downward bias for variance the intercept for the ML method has narrower 95% confidence intervals compared to the REML method for the model fitting a random slope.

Part F Interpret these findings in 1-2 sentences.

Both models have approximately the same output. The intercept of 9.10 is the expected log of the average titer at age 0, which is significant in both models with p-values <0.001. The "Age At Draw" coefficient is the expected change in titer as age progresses each year, with the coefficient meaning there is an expected increase of 0.005 in log of titer each year. This coefficient is insignificant with a p-value >0.05.

Part G What test was used to generate the p-value and what were the denominator DF?

For both the REML and ML methods the Satterthwaite's method was used to generate the p-value. For the REML method the degrees of freedom for the intercept was 76.57 (closer to the number of subjects), and for the "Age At Draw" term it was 425.3 (Closer to the number of total observations). For the ML method the degrees of freedom for the intercept was 79.49 (closer to the number of subjects), and for the "Age at Draw" term it was 426.6 (closer to the number of observations).

Part H Use the anova package to test the significance of the slope on age using a Satterwaithe and a Kenward-Roger DF adjustment. Present and compare the denominator DF. Provide a 1-2 sentence intuitive answer as to why they differ (or are similar). Were there any differences in your study conclusion using different tests?

Table 1: DDF for Different Methods

Method	DDF	P-Value
Satterthwaite	425.274	0.147
Kenward-Roger	425.412	0.147

The DDF are nearly identical for the slope between the Satterthwaite and Kenward-Roger methods. The sample size may be sufficient to overcome the unbalanced data (where the Satterthwaite tends to performs worse). The p-values were the same to the third decimal place, meaning the same conclusion, that the slope is not significant in the random intercept model.

Part I What were the estimates of G_i, R_i, and V_i for this model using both approaches? Interpret the different sources of variation. What estimates will be more biased (ML and REML)? Why?

REML

G_i and R_i are both from the model summary. V_i is the summation of these two components:

- $G_i = 2.0834$
- $R_i = \sigma_e = 0.1154$; This estimate will be in a matrix with the format $\sigma_e I_n$
- $V_i = 2.1988$; This estimate will be in a matrix with dimensions (nx1).

ML

G_i and R_i are both from the model summary. V_i is the summation of these two components:

- $G_i = 2.0299$
- $R_i = \sigma_e = 0.1151$; This estimate will be in a matrix with the format $\sigma_e I_n$
- $V_i = 2.1450$; This estimate will be in a matrix with dimensions (nx1).

The estimates from the ML method will be more biased (downward).

Part J Compute the ICC and interpret. What source of variation is largest, within or between subject variation? G_i is the between subject variation, R_I is the within subject variation, and V_i is the total variation for a specific subject.

REML

$$ICC = \frac{\sigma_{b0}}{\sigma_{b0} + \sigma_e} = \frac{2.0834}{2.0834 + 0.1154} = 0.948$$

In the REML method there is more between subject variation compared to within subject variation.

ML

$$ICC = \frac{\sigma_{b0}}{\sigma_{b0} + \sigma_e} = \frac{2.0299}{2.0299 + 0.1151} = 0.946$$

In the ML method there is more between subject variation compared to within subject variation.

Part K Using the ML estimates, create the EBLUP and the fitted estimates. The individual fitted values of the outcome are a weighted combination of two quantities. What are these two quantities? What determines the weight of the individual estimate towards each of the quantities?

Question 3

Part A Write out the random intercept model for these data in subject level notation including indices for matrices.

$$Y_i = X_i\beta + Z_i b_i + E_i$$

Where each matrix has the following dimensions:

- $Y_i : (n_i \times 1)$; n_i refers to the number of time points (age). Outcome is the log of average titer.

- $X_i : (n_i \times p)$; n_i refers to the number of time points(age) and p is the number of covariates plus the intercept. The only covariates are the number of time points (age).
- $\beta : (p \times 1)$; p refers to the number of covariates (measurements at a certain age plus the intercept).
- $Z_i : (n_i \times q)$; n_i refers to the number of time points (age). q refers to the number of covariates for the random effects. In this case there is a random intercept and a random slope, so $q = 2$.
- $b_i : (q \times 1)$; q refers to the number of covariates for the random effects. In this case there is a random intercept and random slope, so $q = 2$.
- $E_i : (n_i \times 1)$; n_i refers to the number of time points (age). Each individual outcome has its own error term.

Part B What is the format for G_i and R_i that will be assumed in the lmer R function when you fit the random intercept model? Please write the format for these two matrices and interpret. Note we want the random effects to have an unstructured G_i matrix.

The variability of the random effect will be a matrix with $(q_i \times q_i)$ dimensions. Because there is a random intercept and a random slope the G_i matrix will have dimensions of (2×2).

$$G_i = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{bmatrix} = \begin{bmatrix} \sigma_0^2 & \sigma_{01}^2 \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}$$

It is usually not possible to estimate both G and an unstructured R_i, meaning that $R_i = \sigma_e^2 I_n$ (an independence matrix). The dimensions of R_i are $(n_i \times n_i)$, which in this case means a square matrix where n_i is the number of time points (measurements at different ages) for each subject. Because different subjects have different time points the dimensions of the R_i matrix will change between subjects.

Part C Fit a random intercept and slope model for these data using the lmer R function with REML and ML estimation approaches. Provide code and output.

```
### START QUESTION 3 CODE ###

## START PART C CODE ##

# Fitting a lmer model with a random intercept and slope  using the REML method.
measles_ri_rs_reml <- lmer(Avg_Titer_Log ~ Age_At_Draw + (1 + Age_At_Draw|ID),
                           data = measles_raw, REML = TRUE)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0294766 (tol = 0.002, component 1)
```

```
summary(measles_ri_rs_reml)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Avg_Titer_Log ~ Age_At_Draw + (1 + Age_At_Draw | ID)
##    Data: measles_raw
##
## REML criterion at convergence: 364.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.5714 -0.2585 -0.0076  0.2536  7.3816
##
## Random effects:
```

```
##  Groups   Name        Variance  Std.Dev. Corr
##  ID       (Intercept) 11.150144 3.33918
##           Age_At_Draw  0.006077 0.07795  -0.91
##  Residual              0.060553 0.24608
## Number of obs: 454, groups:  ID, 40
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  8.35503    0.55872 37.91618  14.954   <2e-16 ***
## Age_At_Draw  0.02600    0.01297 34.59215   2.005   0.0528 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## Age_At_Draw -0.917
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.0294766 (tol = 0.002, component 1)
```

```
# Fitting a lmer model with a random intercept and slope using the ML method
measles_ri_rs_ml <- lmer(Avg_Titer_Log ~ Age_At_Draw + (1 + Age_At_Draw|ID),
                         data = measles_raw, REML = FALSE)

summary(measles_ri_rs_ml)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: Avg_Titer_Log ~ Age_At_Draw + (1 + Age_At_Draw | ID)
##    Data: measles_raw
##
##      AIC      BIC   logLik deviance df.resid
##    368.7    393.4   -178.4    356.7      448
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.5893 -0.2578 -0.0075  0.2538  7.3893
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  ID       (Intercept) 10.891284 3.30019
##           Age_At_Draw  0.005917 0.07692  -0.91
##  Residual              0.060546 0.24606
## Number of obs: 454, groups:  ID, 40
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  8.35606    0.55266 38.51190  15.120   <2e-16 ***
## Age_At_Draw  0.02597    0.01281 35.08650   2.027   0.0503 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## Age_At_Draw -0.917
```

```
## FINISH QUESTION 3 PART C CODE ##
```

Part D What are the fixed effects estimates with 95%CI and p-values for both fits? How do they compare? Explain.

Term	Estimate	95% Conf.Low	95% Conf.High	P-Value
REML				
Intercept	8.35503	7.22388	9.48618	$<$ 2e-16
Age At Draw	0.02600	-0.00034	0.05234	0.052831
ML				
Intercept	8.35606	7.23775	9.47437	$<$ 2e-16
Age At Draw	0.02597	-0.00003	0.05197	0.050278

The REML method had a slightly lower intercept and a slightly higher "Age At Draw" coefficient. Intercepts were significant in both models and "Age At Draw" coefficients were insignificant in both models. Additionally the ML model gave a warning concerning convergence. Because the variance estimate of the ML method tends to be biased downwards this is why the 95% confidence intervals of the ML method are slightly narrower here after fitting both a random slope and a random intercept.

Part E Interpret these finding in 1-2 sentences.

Both models have approximately the same outcomes. The intercept means that subjects at age 0 are expected to have a log of average titer of 8.36, which is significant with a p-value $<$0.001. The "Age at Draw" coefficient means that for every year subjects are expected to increase their log of average titer by 0.026. This estimate is insignificant in both models, with p-values $>$ 0.05.

Part F What test was used to generate the p-value and what were the denominator DF?

For both the REML and ML methods the Satterthwaite's method was used to generate the p-value. For the REML method the degrees of freedom for the intercept was 37.92 (closer to the number of subjects), and for the "Age At Draw" coefficient it was 34.59 (closer to the number of subjects). For the ML method the degrees of freedom for the intercept was 38.51 (closer to the number of subjects), and for the "Age At Draw" coefficient it was 35.09 (closer to the number of subjects).

Part G How did the denominator DF in the test statistic differ between the random intercept and random slope model? Explain what is happening here (if anything).

For the test statistics denominator DF in the random intercept only model was closer to the number of subjects. Once a random slope was added to the model the test statistic denominator DF was closer to the number of subjects. This makes sense as fitting a random slope for each subjects uses up DF, and inferences on slope are collapsed closer to the number of subjects compared to the number of observations after random slopes are fitted.

Part H Use the anova package to test the significance of the slope on age using a Satterthwaite and a Kenward-Roger DF adjustment for the REML model. Present and compare the denominator DF. Provide a 1-2 sentence intuitive answer as to why they differ (or are similar). Were there any differences in your conclusions between these different testing approaches.

Table 2: Random Intercept-Slope REML DF

Method	DDF	P-Value
Satterthwaite	34.592	0.053
Kenward-Roger	38.429	0.052

The DDF for the Satterthwaite method is less than the Kenward-Roger method. Because a random slope was fitted this grouped data for the slope from being individual observations as it was in the random intercept only model to being observations within subjects as age progressed. This reduction accounts for both the DDF being closer to the number of subjects instead of the number of observations, as well as the difference in DDF observed between the two methods. Due to unbalanced data between subjects the Kenward-Roger method should be chosen. The p-values are very close however, and neither is statistically significant.

Part I Were there any differences between tests and denominator DF between the random intercept and random slope models? Explain.

Between the random intercept and random intercept/slope models neither p-values are significant. However, the p-values from the random intercept/slope models are much closer to the level of significance due to accounting for between subject variation in slopes. Between the random intercept and random intercept/slope models the degrees of freedom in the random intercept model is much higher and much closer to the number of observations. After accounting for between subject differences in slopes with the random slope/intercept model the DF are closer to the number of subjects.

Part J What were the estimates of G_i, R_i, and V_i for this model using both approaches? Interpret the different sources of variation. What estimates will be more biased (ML and REML)? Why?

All values were extracted from the model outputs.

REML

- G_i:

$$G_i = \begin{bmatrix} \sigma^2_{b0} & \sigma_{b,01} \\ \sigma_{b,01} & \sigma^2_{b1} \end{bmatrix} = \begin{bmatrix} 11.15 & -0.91 \\ -0.91 & 0.0061 \end{bmatrix}$$

- R_i: $\sigma_e = 0.0606$; R_i is a matrix with form $\sigma_e I_n$
-