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_data, 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_data
##
## 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_data, 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_data
##
##      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 Random Intercept - 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 (nxn).

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

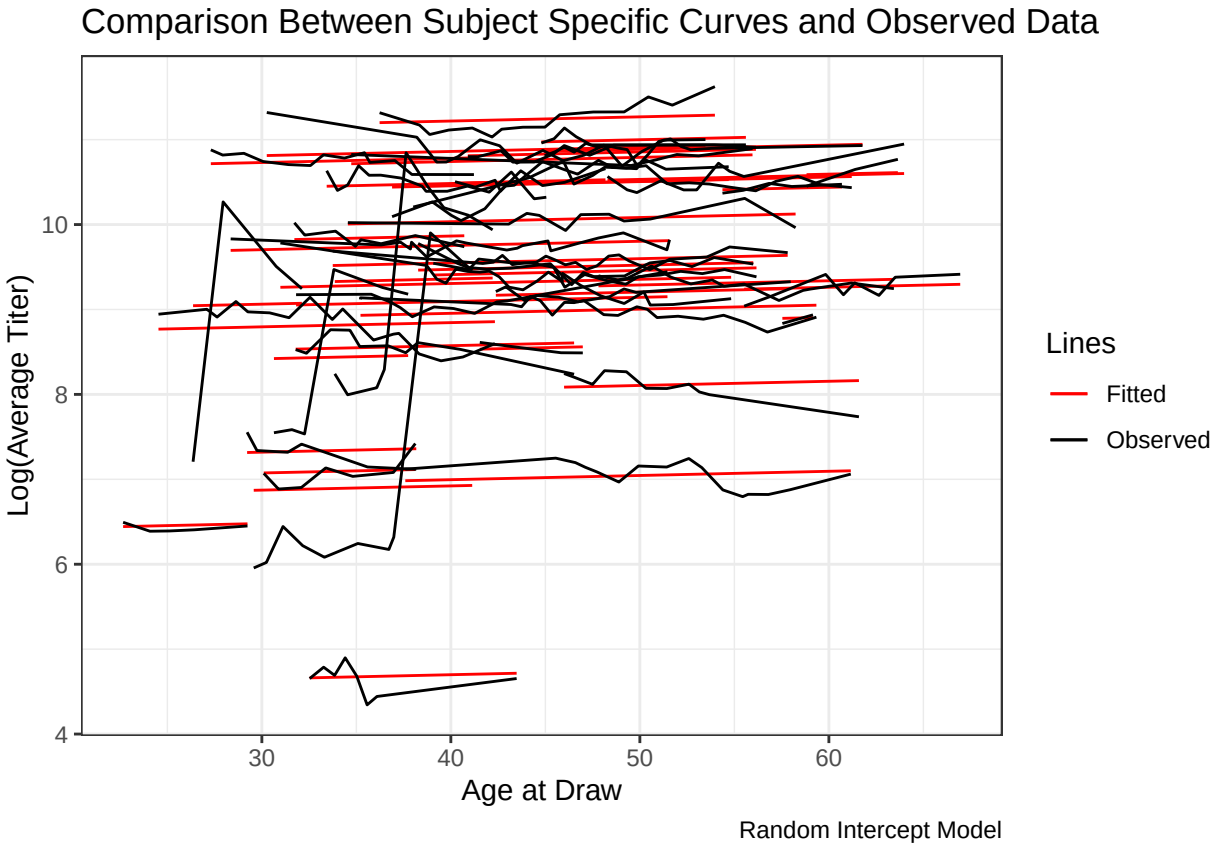
$$ICC = \frac{\sigma_{b0}}{\sigma_{b0} + \sigma_e} = \frac{2.0299}{2.0299 + 0.1151} = 0.946$$

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?

The individual fitted values are a combination of subject specific data (random effects) and group averaged data. The weight of individual estimates is the difference between the population average profile and the observed data.

Part L Make a plot of the estimated individual specific curves compared to the observed data. How well do you think this model is capturing the the variation in the data and the age trends in the data? Provide an interpretation in a few sentences.



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,
                        control = lmerControl(optimizer = "bobyqa"))

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
## Control: lmerControl(optimizer = "bobyqa")
##
## REML criterion at convergence: 364.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.5705 -0.2587 -0.0076  0.2537  7.3821
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  ID       (Intercept) 11.225751 3.3505
##           Age_At_Draw  0.006115 0.0782   -0.91
##  Residual              0.060513 0.2460
## Number of obs: 454, groups:  ID, 40
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  8.35466    0.56044 37.59941   14.91   <2e-16 ***
## Age_At_Draw  0.02602    0.01301 34.34728    2.00   0.0534 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## Age_At_Draw -0.917
```

```r
# 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
```

```r
## 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	8.35466	7.21970	9.48961	< 2e-16
Age At Draw	0.02602	-0.00040	0.05244	0.053408
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	34.347	0.053
Kenward-Roger	38.433	0.053

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 $\sigma_e I_n$
- V_i:

$$V_i = \begin{bmatrix} 1 & age_{i1} \\ 1 & age_{i2} \\ \vdots & \vdots \\ 1 & age_{in} \end{bmatrix} \begin{bmatrix} 11.15 & -0.91 \\ -0.91 & 0.0061 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ age_{i1} & age_{i2} & \cdots & age_{in} \end{bmatrix} + 0.0606 I_n$$

ML

- G_i

$$G_i = \begin{bmatrix} \sigma^2_{b0} & \sigma_{b,01} \\ \sigma_{b,01} & \sigma^2_{b1} \end{bmatrix} = \begin{bmatrix} 10.89 & -0.91 \\ -0.91 & 0.0059 \end{bmatrix}$$

* R_i: $\sigma_e = 0.0605$; R_i is a matrix with form $\sigma_e I_n$ * V_i:

$$V_i = \begin{bmatrix} 1 & age_{i1} \\ 1 & age_{i2} \\ \vdots & \vdots \\ 1 & age_{in} \end{bmatrix} \begin{bmatrix} 10.89 & -0.91 \\ -0.91 & 0.0059 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ age_{i1} & age_{i2} & \cdots & age_{in} \end{bmatrix} + 0.0605 I_n$$

Part K Using the ML estimates, compute the ICC for the slope and interpret. What percentage of variation is due to the random effect on slope.

$$ICC(Slope) = \frac{\sigma^2_{b1}}{\sigma^2_{b0} + \sigma^2_{b1} + \sigma^2_e} = \frac{0.005917}{10.891284 + 0.005917 + 0.060546} = 5.3998 \times 10^{-4}$$

The ICC for slope is very small meaning that almost none of the variance is due to the random effect on slope. This means that there is not a lot of variance in the slopes between subjects.

$$ICC(Intercept) = \frac{\sigma^2_{b0}}{\sigma^2_{b0} + \sigma^2_{b1} + \sigma^2_e} = \frac{10.891284}{10.891284 + 0.005917 + 0.060546} = 0.994$$

The ICC for the intercept accounts for nearly of the variance in the model. This means that there is a lot of variance in the intercepts between subjects. In other words, subjects start at vastly different average titer levels.

Part L Using the ML estimates, create the EBLUP and the fitted estimates. There is nothing to turn in for this question but the code.

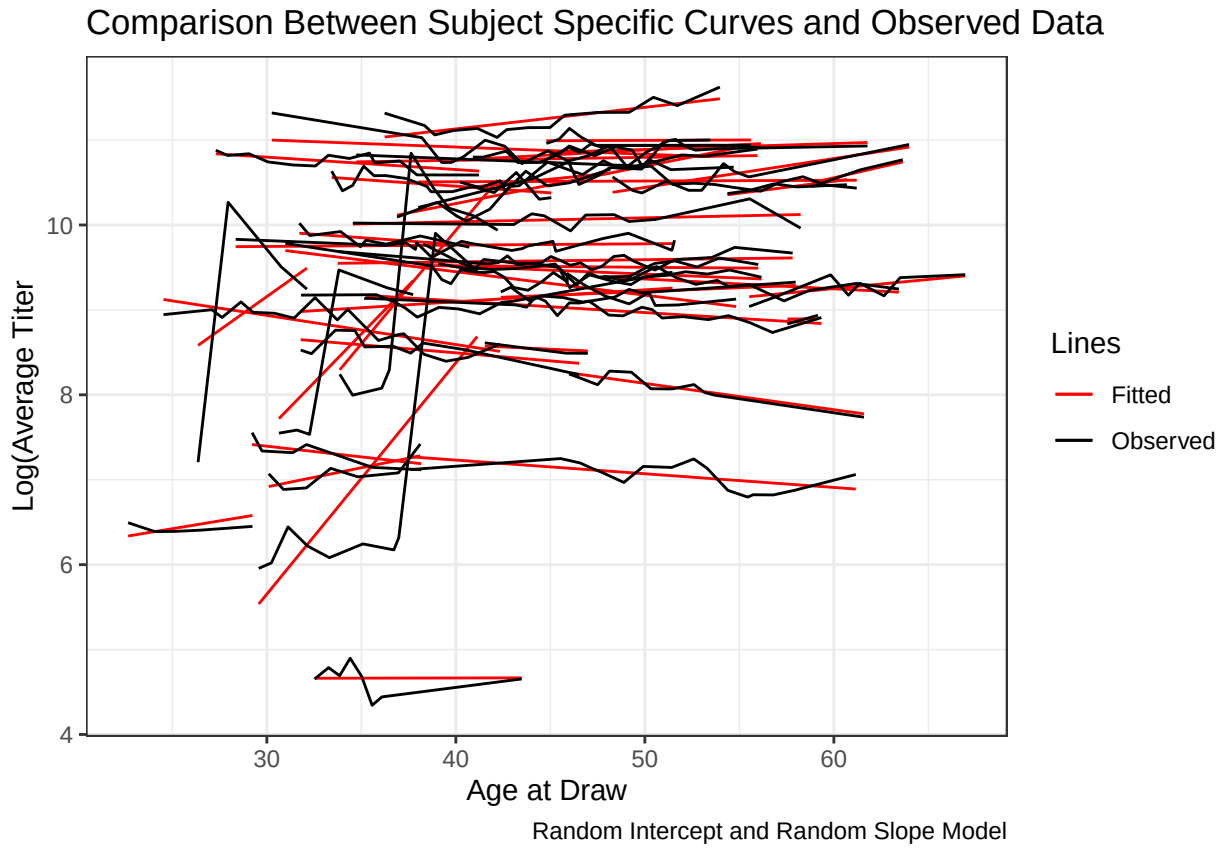
```
## START QUESTION 3 PART L CODE ##

# Calculating the estimates. Converting to data.frame to extract the
ri_rs_eblups <- ranef(measles_ri_rs_ml,condVar = TRUE)

# Calculating the fitted values.
ri_rs_fitted <- fitted(measles_ri_rs_ml)

## FINISH QUESTION 3 PART L CODE ##
```

Part M Make a plot of the estimated individual specific curves compared to the observed data. How well do you think this model is capturing the variation in the data and the age trends in the data? Provide an interpretation in a few sentences.



For each individual subject the random intercept/slope model more closely models the trend between age and the log of average titer. Most subjects have a roughly linear trend with time and the random slopes model these well, although a few subjects have trends which are not linear.

Part N Base on the fits do you think the random slope is a critical additional component to add to the model? Explain.

Based on the ICC adding a random slope does not significantly account for variance within the model. From the plot of fitted values the fitted lines do better track with subject trends, but subjects vary enough along the y-axis that doing this does not significantly fit the model better.

Part O The investigator is leery of these mixed effects models. How would you show him/her what the random intercept and random slope model did for estimation compared to just fitting a bunch of regression models, one for each person?

Using a random intercept/slope model allows for each subject to be modeled separately, which in effect provides the same inferences on subjects that fitting multiple regression models for each subject would. The random intercept/slope model also provides the average trends, which are what will be used for making inferences on people who have had measles vaccines.

Part 2

Question 3

Part A Use a Poisson GLM (i.e. Poisson regression) to estimate the association between condition (group) and number of breakfast servings, adjusting for sex and weight.

Table 3: Condition - Poisson Model - Exponentiated

Term	Estimate	Std.Error	95% Conf.Low	95% Conf.High	P-Value
(Intercept)	2.7127	0.1988	1.8301	3.9913	5.1438e-07
Cond	2.3739	0.1350	1.8384	3.1234	1.5045e-10
Sex1	0.8197	0.0822	0.6971	0.9624	0.015612
Wt1	1.2181	0.1323	0.9372	1.5745	0.135932

1.) $Y_i \sim Pois(\lambda_i)$; $E[Y_i] = Var[Y_i] = \lambda_i$

2.) $\eta_i = \beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Weight_i$

3.) $\eta_i = log(\lambda_i)$

The data for this analysis came from week 1 (baseline) measurements for cereal consumption for Kid1. From the Poisson regression children in the experimental group were expected to have a rate of consumption of breakfast cereal servings 2.37 (95% CI:1.84,3.12) times greater than that of the those who did not receive anything special when controlling for sex and weight. This is a significant result with a p-value of <0.001.

Part B Repeat (a) allowing for overdispersion by using quasilikelihood with the Poisson GLM.

Table 4: Condition - QuasiPoisson Model - Exponentiated

Term	Estimate	Std.Error	95% Conf.Low	95% Conf.High	P-Value
(Intercept)	2.7127	0.3411	1.3728	5.2435	0.00429612
Cond	2.3739	0.2316	1.5463	3.8517	0.00032294
Sex1	0.8197	0.1411	0.6201	1.0790	0.16209613
Wt1	1.2181	0.2271	0.7742	1.8861	0.38706649

1.) $Y_i \sim Pois(\lambda_i)$; $E[Y_i] = \lambda_i$; $V[Y_i] = \phi\lambda_i$

2.) $\eta_i = \beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Weight_i$

3.) $\eta_i = log(\lambda_i)$

The data for this analysis came from week 1 (baseline) measurements for cereal consumption for Kid1. From the Quasi-Poisson regression children in the experimental group were expected to have a rate of consumption of breakfast cereal servings 2.37 (95% CI:1.546,3.852) times greater than that of the those who did not receive anything special when controlling for sex and weight. This is a significant result with a p-value of <0.001.

Part C Repeat (a) allowing for overdispersion by adding a random normal error to the linear predictor in the Poisson GLM and using maximum likelihood estimation.

Table 5: Condition - Poisson Random Intercept Model - Exponentiated

Term	Estimate	Std.Error	95% Conf.Low	95% Conf.High	P-Value
(Intercept)	2.4151	0.7830	1.2793	4.5594	0.0065347
Cond	2.3517	0.4634	1.5983	3.4602	1.4262e-05
Sex1	0.8577	0.1229	0.6477	1.1359	0.2842532
Wt1	1.1779	0.2781	0.7415	1.8711	0.4880850

1.) $Y_i|\epsilon_i \sim Pois(\lambda_i)$; $E[Y_i] = \lambda_i$; $Var[Y_i] = \lambda_i + (e^{\sigma_\epsilon^2} - 1)\lambda_i^2$

2.) $\eta_i = \beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Weight_i + \epsilon_i$; $\epsilon_i \sim N(0, \sigma_{\epsilon i}^2)$

3.) $\eta_i = log(\lambda_i)$

The data for this analysis came from week 1 (baseline) measurements for cereal consumption for Kid1. From the Poisson with normal random error regression children in the experimental group were expected to have a rate of consumption of breakfast cereal servings 2.35 (95% CI:1.60,3.46) times greater than that of the those who did not receive anything special when controlling for sex and weight. This is a significant result with a p-value of <0.001.

Part D Repeat (a) allowing for overdispersion by using a Negative Binomial GLM estimated with maximum likelihood.

Table 6: Condition - Negative Binomial Model - Exponentiated

Term	Estimate	Std.Error	95% Conf.Low	95% Conf.High	P-Value
(Intercept)	3.0208	0.3229	1.5868	5.7083	0.00061734
Cond	2.3569	0.1955	1.6025	3.4695	1.1558e-05
Sex1	0.8428	0.1437	0.6381	1.1136	0.23375812
Wt1	1.1046	0.2340	0.6895	1.7809	0.67085322

1.) $Y_i \sim NB(k_i, \lambda_i)$; $E[Y_i] = \frac{k_i(1-\lambda_i)}{\lambda_i}$; $V[Y_i] = \frac{k_i(1-\lambda_i)}{\lambda_i^2}$

2.) $\eta_i = \beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Weight_i$

3.) $\eta_i = log(\frac{\lambda_i}{\lambda_i + k_i})$

The data for this analysis came from week 1 (baseline) measurements for cereal consumption for Kid1. From the Negative Binomial regression children in the experimental group were expected to have a rate of consumption of breakfast cereal servings 2.36 (95% CI:1.60,3.47) times greater than that of the those who did not receive anything special when controlling for sex and weight. This is a significant result with a p-value of <0.001.

Question 4

Part A Make a table summarizing the results of the previous models.

	"Poisson Regression"	"Poisson QL"	"Poisson + Normal error"	"NB NLMIXED"
Intercept	0.998(0.199)	0.998(0.3411)	0.882(0.3242)	1.106(0.323)
Cond	2.374(1.838,3.123)	2.374(1.546,3.852)	2.352(1.598,3.460)	2.357(1.603,3.470)

	"Poisson Regression"	"Poisson QL"	"Poisson + Normal error"	"NB NLMIXED"
Sex	0.820(0.697,0.962)	0.820(0.620,1.079)	0.858(0.648,1.136)	0.843(0.638,1.114)
Wt	1.218(0.937,1.575)	1.218(0.774,1.886)	1.178(0.742,1.871)	1.105(0.690,1.781)
Other	NA	2.9447	0.5429	1

Part B Write a short paragraph summarizing the results of comparing conditions, e.g. which condition gives higher consumption and by how much.

The data for this analysis came from week 1 (baseline) measurements for cereal consumption for Kid1. From the Quasi-Poisson regression children in the experimental group were expected to have a rate of consumption of breakfast cereal servings 2.37 (95% CI:1.546,3.852) times greater than that of the those who did not receive anything special when controlling for sex and weight. This is a significant result with a p-value of <0.001. The Quasi-Poisson regression model was chosen to account for characteristics of the data.

Part c Write a short paragraph comparing the model estimated, e.g. differences between parameter estimates across models.

Both the Poisson and Quasi-Poisson models have similar estimates for all coefficients. The 95% CI is always wider for the Quasi-Poisson model, which is due to the Quasi-Poisson model attempting to address overdispersion. This is illustrated in the "Other" row for the "Poisson QL" column, where the scale parameter is 2.9447. For the Poisson + Normal error model the coefficient is greater for Sex, and less for condition and weight. For the NB NLMIXED model the intercept is larger compared to the other models. For the condition and sex coefficients in the NB NLMIXED model the value is between the same coefficient values in other models. For the wt coefficient in the NB NLMIXED model the value is lower that any other model.

Question 5

Part A Fit a logistic regression of these data and summarize the relationship between the dues increase and not renewing. Interpret the findings including a 95% CI and p-value. Write 1-2 complete sentences describing the findings.

Table 8: Nonrenewal Logistic Regression - Exponentiated

Term	Estimate	Std.Error	95% Conf.Low	95% Conf.High	P-Value
(Intercept)	0.0082	2.6558	0.0000	1.1039	0.070261
dueincrease	1.1332	0.0668	1.0023	1.3097	0.060985

From the logistic regression model for each increase in $1 of dues there is a 1.13 times greater rate of nonrenewal (95% CI:1.0023,1.3097) with an insignificant p-value of 0.061.

Part B Compute the estimated probability that association members will not renew their membership if the dues are increased by $40.

If membership dues are increased by $40 55% of subjects are estimated to not renew.

Part C Estimate the amount of dues increase for which 75% of the members are not expected to renew their association membership.

The estimated amount of dues increase for which 75% of members are not expected to renew their association membership is $47.

Part D Is the default p-value and test a Wald based test or a likelihood based test?

The p-value is a Wald based test (based on the Z-Score).

Part E Fit a quasi binomial model to these data and assess whether there is any evidence of overdispersion. Write a few sentences comparing the parameter estimates and SE's (and p-values) from this model and the regular logistic regression model (without overdispersion).

Table 9: Nonrenewal Quasi-Binomial Regression - Exponentiated

Term	Estimate	Std.Error	95% Conf.Low	95% Conf.High	P-Value
(Intercept)	0.0082	2.7537	0.0000	1.3136	0.091802
dueincrease	1.1332	0.0692	0.9979	1.3176	0.081523

Taking the ratio of deviance and residual degrees of freedom returns 1.34. This is below the general threshold of 1.5, meaning that over dispersion is likely not an issue for this data set.

Comparing the two models:

- Intercept- Logistic(Exp. Est.:0.0082; SD:2.6558); Quasi-Binomial(Exp. Est.:0.0082; SD:2.7537). The estimates are the same, but the standard error is larger in the Quasi-Binomial model. This is likely due to the Quasi-Binomial adjusting for the small amount of overdispersion that is present in the model.
- Due Increase - Logistic(Exp. Est.:1.1332; SD:0.0668); Quasi-Binomial(Exp. Est.:1.1332; SD:0.0692). Thes estimates are the same, but the standard error is larger in the Quasi-Binomial model. As above, this may be due to the Quasi-Binomial adjusting for the small amount of overdispersion that is present in the model.