Longitudinal Data Analysis 2022: Exercise bundle

October 21, 2022

1 ANOVA

Question A1. This question is about factors involved in the school data set.

(a) Mention all factors involved in the case study data.

(b) Identify if these factors are nested or crossed.

(c) Determine all possible interactions between factors.

(d) Determine which factors should be treated as fixed and random.

Question A2. Consider a one way ANOVA model with a random effect (see model definition on slide 14). Answer the following theoretical questions:

(a) Demonstrate that the correlation between two different observations on the same unit (i.e. between y_{ir} and y_{is}) is equal to the ICC.

(b) Demonstrate that the sums of squares for the within group variability and between group variability add up to the total sums of squares.

(c) Demonstrate that the variance of the mean of all observations ($\bar{y}_{\cdot\cdot}$) is equal to: $\sum_{i=1}^{m} n_i^2 \sigma_G^2 / n^2 + \sigma_E^2 / n$.

(d) Demonstrate the following:

$$\begin{aligned}
\text{COV}(y_{ij}, \bar{y}_{i\cdot}) &= \sigma_G^2 + \frac{\sigma_E^2}{n_i}, \\
\text{COV}(\bar{y}_{i\cdot}, \bar{y}_{\cdot\cdot}) &= \frac{n_i \sigma_G^2 + \sigma_E^2}{n}.
\end{aligned}$$

(e) Show that the expected mean squares are given by:

$$\begin{aligned}
\mathbb{E}[MS_W] &= \sigma_E^2, \\
\mathbb{E}[MS_B] &= \frac{n - \sum_{i=1}^{m} n_i^2 / n}{m - 1} \sigma_G^2 + \sigma_E^2.
\end{aligned}$$

Question A3. Using the school data, perform a one-way ANOVA on the difference in language test scores for all children with class as a random effect.

(a) Estimate the model parameters.

(b) Obtain the confidence intervals.

(c) Calculate the intraclass correlation coefficient and a 95% confidence interval.

(d) Calculate the total variability and a 95% confidence interval.

Question A4. Using the school data, perform a one-way ANOVA on the difference in language test scores for all children with class as a random effect.

(a) Investigate the BLUPs for each class:

(i) Study the most extreme BLUPs and averages.

(ii) Make a plot of the predictions against the averages.

(b) Investigate the conditional and marginal residuals. Do you think that the conditions of normality are satsified?

Question A5. In this question we investigate Maximum Likelihood estimation for one-way ANOVA models.

(a) Derive the likelihood equations.

(b) Determine the ML (maximum likelihood) solutions when the data is balanced.

(c) Determine the ML estimators when the data is balanced.

(d) Determine the Fisher information matrix.

(e) Determine the inverse Fisher information matrix.

For the next question use the school data. Fit the model in A3 with ML estimation.

(f) Estimate the model parameters

(g) Obtain the confidence intervals

(h) Calculate the intraclass correlation coefficient and a 95% confidence interval.

(i) Calculate the total variability and a 95% confidence interval.

Question A6. In this question we investigate Restricted Maximum Likelhiood estimation for one-way ANOVA models

(a) Derive the likelihood function.

(b) Derive the likelihood equations.

(c) Determine the REML (maximum likelihood) solutions when the data is balanced.

(d) Determine the REML estimators when the data is balanced.

For the next question use the school data. Fit the model in A3 with REML estimation.

(e) Estimate the model parameters

(f) Obtain the confidence intervals

(g) Calculate the intraclass correlation coefficient and a 95% confidence interval.

(h) Calculate the total variability and a 95% confidence interval.

Question A7. Fit a two-way nested mixed effects model, with ANOVA, for the difference in language score, with `COMBI` as fixed effect and `CLASS` as random effect nested within `COMBI`.

(a) Investigate the differences between TYPE1 and TYPE3 estimation.

(b) Report the parameter estimates and their 95% confidence intervals.

(c) Report the total variability and intraclass correlation coefficient with their 95% confidence intervals.

Fit the same two-way nested mixed effects model for the difference in language score using the ML estimator.

(d) Investigate heteroscedasticity for between class variability for the two class types.

(e) Report the relevant information.

Question A8. Consider a balanced crossed mixed effects model (without interactions), with `SEX` and `CLASS` as main factors.

(a) Determine the mean squares.

(b) Determine the expected mean squares.

(c) Determine the estimators for all model parameters.

Fit the two-way crossed ANOVA model for the difference in language score using TYPE1 estimation.

(d) Report the parameters and their 95% confidence intervals.

(e) Calculate the total variability and the intraclass correlation coefficients with their 95% confidence intervals.

(f) Investigate the conditional and marginal residuals

(g) Answer questions (d)-(f) for the three other estimation methods: TYPE3, ML, REML and discuss the differences.

Question A9. Consider a balanced three-way ANOVA model with factors S,T and C(T), with S fixed and T and C(T) random:

(a) Determine the mean squares

(b) Determine the expected mean squares

(c) Determine the estimators for all model parameters

Consider a balanced three-way ANOVA model with factors S,T and C(T), with T fixed and S and C(T) random:

(a) Determine the mean squares

(b) Determine the expected mean squares

(c) Determine the estimators for all model parameters

Question A10. Fit a three-way ANOVA model for the difference in language score with sex, type of class, and class within type of class as factors, using the TYPE1 estimation method.

(a) Report the parameters and their 95% confidence intervals.

(b) Calculate the total variability and the intraclass correlation coefficients with their 95% confidence intervals.

(c) Investigate the conditional and marginal residuals.

(d) Answer questions $(a) - (c)$ for the three other estimation methods: TYPE3, ML, REML and discuss the differences.

Question A11. Consider the following three-way ANOVA model with factors S(fixed), T(fixed), and $C(T)$(random), but with no interactions

(a) Determine the mean squares

(b) Determine the expected mean squares

(c) Determine the estimators for all model parameters

2 Generalized Linear Models

Question L1. Consider the following three-way ANOVA model:

$$y_{hijk} = \mu + \alpha_h + \beta_i + (\alpha\beta)_{hi} + a_{j(i)} + (\alpha a)_{hj(i)} + e_{hijk}$$

with,

- μ the overall mean,
- α_h the effect of sex ($\alpha_2 = 0$),
- β_i the effect of class type ($\beta_2 = 0$),
- $(\alpha\beta)_{hi}$ the interaction effect of sex and class type ($(\alpha\beta)_{11} \neq 0$),
- $a_{j(i)} \sim N\left(0, \sigma^2_{C(T)}\right)$ the effect of class,
- $(\alpha a)_{hj(i)} \sim N\left(0, \sigma^2_{SC(T)}\right)$ the interaction effect of sex and class with type of class,
- $e_{hijk} \sim N\left(0, \sigma^2_R\right)$ the residual

(a) Determine the $\mathbf{X}_i$ and $\mathbf{Z}_i$ matrix for unit i being class.

(b) Determine the $\mathbf{V}_i$ matrix.

Question L2. Fit the following mixed effects model:

$$y_{hijk} = \mu + \alpha_h + \beta_i + (\alpha\beta)_{hi} + b_{hj(i)} + e_{hijk}$$

with,

- μ the overall mean,
- α_h the effect of sex ($\alpha_2 = 0$),
- β_i the effect of class type ($\beta_2 = 0$),
- $(\alpha\beta)_{hi}$ the interaction effect of sex and class type ($(\alpha\beta)_{11} \neq 0$),

- $b_{hj(i)}$ the random effect of sex h within class i for type of class j

$$\left(b_{11(i)}, b_{21(i)}, b_{12(i)}, b_{22(i)}\right)^T \sim N\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{pmatrix}\right),$$

where $\boldsymbol{G}_{jj} = \begin{pmatrix} \sigma_1^2(j) & \rho_j\sigma_1(j)\sigma_2(j) \\ \rho_j\sigma_1(j)\sigma_2(j) & \sigma_2^2(j) \end{pmatrix}$ and $\boldsymbol{G}_{12} = \boldsymbol{G}_{21} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$

- $e_{hijk} \sim N\left(0, \sigma_R^2\right)$ the residual

Question L3. Assuming the subject-specific model with linear time profiles

(a) Formulate the correlation for two repeated observations at time t and $(1+c)t$.

(b) Formulate 95% reference (prediction) limits for the linear time profiles.

Using the outcome AUC of the pain data (`area`) and fit the subject-specific model with linear time profile and treatment.

(c) Report the model parameters with their 95% confidence intervals.

(d) Do you believe that treatment has an effect on the outcome?

(e) Do you believe that time has an effect on the outcome?

(f) Estimate the correlation between observations at 135 and 150.

(g) Estimate the 95% confidence and reference limits for the three treatments.

Question L4. Consider the growth data on child weight with logarithmic transformed weight as the outcome. Fit a fractional polynomial with $p_1 = 1$ and $p_2 = -0.5$ assuming a separate average FP for sexes, and that the three coefficients have a multivariate normal distribution with i.i.d. normal residuals.

(a) Are the average FPS for sexes different?

(b) Does the model predict the observations?

(c) Do the residuals show better or worse patterns than the residuals in the course slides?

(d) Calculate reference limits for the weight growth of children for both sexes seperately.

Question L5. Consider the AUC of the pain data, fit a marginal model with categorical time points. Include the effects time, treatment, and the interaction time and treatment. Use an unstructured residual variance.

(a) Report on the p-values of the overall test statistics for fixed effects

(b) Describe the correlation structure between time points.

(c) Report the difference (and its 95%Cl) between treatment G and P and R and P at time point 165. Are they significant?

(d) Report the difference (and its 95%Cl) between the results at time 120 and the average of the three time points 135,150, and 165 for the three treatments. Are they significant?

Question L6. Consider the growth data on child weight, use the logarithmic transformed weight as outcome and fit a marginal model with fractional polynomial (with $p_1 = 1$ and $p_2 = -0.5$). Use separate average FP for sexes, and assume that the structure of the residual variance matrix is unstructured.

(a) Are the average FPS for sexes different?

(b) Calculate reference limits for the weight growth of children for both sexes separately.

(c) Compare these reference limits with the limits of the subject-specific model.

Question L7. Consider the growth data on child weight. Use the logarithmic transformed weight as outcome. Fit a marginal model with fractional polynomial (with $p_1 = 1$ and $p_2 = -0.5$). Use a separate average FP for sexes and assume that the structure of the residual variance matrix is unstructured. Investigate different models which considers: covariance structures for the residuals and the importance of sex in the model. Report which model seems most reasonable.

3 Solutions

Solution A1.

(a) Factor variables are *categorical* variables, and include in this study `child, class, combi, girl, minority, sitters.` (Semi-)continuous variables included in the study (socioeconomic statuses, class sizes, and verbal & performal IQ scores) could be dichotomized to obtain factors. We ignore these here.

(b) Factors are crossed if levels of one factor co-occur with all levels of the other factor. E.g., sex and class are crossed since classes include both girls and boys, and sex measures the same construct across classes. Factors are nested if the levels of one factor co-occur with exactly one level of the other factor. E.g., `class` and `child` are nested, since all children go to exactly one class and children differ across classes. `class` is nested within `combi` as each class is either combi or not, and classes differ across `combi`.Cross-tabulation of the factors can help determine which are nested / crossed.

(c) Interactions are of interest if we hypothesize that there are differential effects for one factor across levels of another factor. Such interaction effects cannot be studied for fully nested factors (e.g. child & class), since we observe outcomes for each child in exactly one class. Interaction effects can be studied for crossed factors. E.g., class & girl, class & minority, etc.

(d) Factors for which we exhaustively observe the levels should be interpreted as fixed. These include `girl, minority, sex, combi` and `sitters`. Factors should be interpreted as random if we do not exhaustively observe levels and are interested in studying the population, rather than the included factor levels at hand. For this study, classes were randomly sampled from randomly sampled schools. Our interest is likely in drawing conclusions about all schools / classes. Hence, `class / school` should be interpreted as random.

Solution A2.

(a) From the definition of the model, $y_{ij} = \mu + a_i + \varepsilon_{ij}$, where a_i are i.i.d. $N(0, \sigma_G^2)$ and ε_{ij} are i.i.d. $N(0, \sigma_E^2)$ random variables and are independent of each other. Therefore by independence of $a_i, \varepsilon_{i,j}$,

$$\text{Var}\,(y_{ij}) = \text{Var}\,(y_{ir}) = \text{Var}(a_i) + \text{Var}(\varepsilon_{ij}) = \sigma_G^2 + \sigma_E^2 \ .$$

On the other hand,

$$\begin{aligned}\text{Cov}\,(y_{ij}, y_{ir}) =& \text{Cov}\,(\mu + a_i + \varepsilon_{ij}, \mu + a_i + \varepsilon_{ir})\\ =& \text{Cov}\,(a_i + \varepsilon_{ij}, a_i + \varepsilon_{ir})\\ =& \text{Var}(a_i) + \text{Cov}(a_i, \varepsilon_{ir}) + \text{Cov}(a_i, \varepsilon_{ij}) + \text{Cov}(\varepsilon_{ir}, \varepsilon_{ij})\\ =& \sigma_G^2.\end{aligned}$$

The last 3 terms produce zero because of the independent structure. Hence,

$$\text{cor}(y_{ij}, y_{ir}) = \frac{\text{Cov}(y_{ij}, y_{ir})}{\sqrt{\text{Var}(y_{ir})\text{Var}(y_{ij})}} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2} \ .$$

(b) We start with the argument in the LHS.

$$\begin{aligned}
\mathrm{SS_T} =&\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{..})^2\\
=&\sum_{i=1}^{m}\sum_{j=1}^{n_i}[(y_{ij}-\overline{y}_{i.})+(\overline{y}_{i.}-\overline{y}_{..})]^2\\
=&\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2+\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\overline{y}_{i.}-\overline{y}_{..})^2+2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})(\overline{y}_{i.}-\overline{y}_{..})\\
=&\sum_{i=1}^{m}\mathrm{SS_{W,i}}+\sum_{i=1}^{m}n_i(\overline{y}_{i.}-\overline{y}_{..})^2+2\sum_{i=1}^{m}(\overline{y}_{i.}-\overline{y}_{..})\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})\\
=&\mathrm{SS_W}+\mathrm{SS_B}\ ,
\end{aligned}$$

since the last term in the penultimate equality produces 0 by the definition of $\overline{y}_{i.}$.

(c) Expanding $\overline{y}_{..}$ from definition,

$$\mathrm{Var}(\overline{y}_{..})=\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}y_{ij}\right)=\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}y_{ij}\right)=\frac{1}{n^2}\mathrm{Var}\left(n\mu+\sum_{i=1}^{m}n_ia_i+\sum_{i=1}^{m}\sum_{j=1}^{n_i}\varepsilon_{ij}\right). \quad (1)$$

Now using the independence structure of the a_i and ε_{ij} and the fact that $\mathrm{Var}(c+X)=\mathrm{Var}(X)$ and $\mathrm{Var}(cX)=c^2\mathrm{Var}(X)$ for all constant c and random variable X, from (1),

$$\mathrm{Var}(\overline{y}_{..})=\frac{1}{n^2}\left[\sum_{i=1}^{m}n_i^2\mathrm{Var}(a_i)+\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathrm{Var}(\varepsilon_{ij})\right]=\sigma_G^2\sum_{i=1}^{m}\frac{n_i^2}{n^2}+\frac{\sigma_E^2}{n}\ .$$

(d) Again we expand the $\overline{y}_{i.}$ in the LHS and obtain,

$$\mathrm{Cov}\,(y_{ij},\overline{y}_{i.})=\mathrm{Cov}\left(y_{ij},\frac{1}{n_i}\sum_{r=1}^{n_i}y_{ir}\right)=\frac{1}{n_i}\left[\mathrm{Var}(y_{ij})+\sum_{\substack{r=1\\r\neq j}}^{n_i}\mathrm{Cov}(y_{ij},y_{ir})\right].$$

From (a), we have the required covariance term. Therefore,

$$\mathrm{Cov}\,(y_{ij},\overline{y}_{i.})=\frac{1}{n_i}\left[\sigma_G^2+\sigma_E^2+\sum_{\substack{r=1\\r\neq j}}^{n_i}\sigma_G^2\right]=\frac{n_i\sigma_G^2+\sigma_E^2}{n_i}=\sigma_G^2+\frac{\sigma_E^2}{n_i}\ .$$

To prove the other part, we keep the $\overline{y}_{i.}$ intact and break the other term.

$$\mathrm{Cov}(\overline{y}_{i.},\overline{y}_{..})=\frac{1}{n}\mathrm{Cov}\left(\overline{y}_{i.},\sum_{k=1}^{m}\sum_{j=1}^{n_k}y_{kj}\right)\ . \quad (2)$$

Now again using independence between $\overline{y}_{i.}$ and y_{kj} for $k\neq i$, from (2), we obtain,

$$\begin{aligned}
\mathrm{Cov}(\overline{y}_{i.},\overline{y}_{..})=&\frac{1}{n}\mathrm{Cov}\left(\overline{y}_{i.},\sum_{j=1}^{n_i}y_{ij}\right)=\frac{n_i}{n}\mathrm{Var}(\overline{y}_{i.})\\
=&\frac{n_i}{n}\mathrm{Var}(\mu+a_i+\overline{\varepsilon}_{i.})\\
=&\frac{n_i}{n}\left[\mathrm{Var}(a_i)+\frac{1}{n_i}\mathrm{Var}(\varepsilon_{ij})\right]\\
=&\frac{n_i\sigma_G^2+\sigma_E^2}{n}\ .
\end{aligned} \quad (3)$$

(e) From the definition of $\mathrm{MS_W}$, we have

$$\mathbb{E}[\mathrm{MS_W}] = \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbb{E}\left(y_{ij}-\overline{y}_{i.}\right)^2 \ .$$

Analysing each of the summands on the RHS,

$$\begin{aligned}\mathbb{E}\left(y_{ij}-\overline{y}_{i.}\right)^2 &= \mathbb{E}\left[(y_{ij}-\mu)-(\overline{y}_{i.}-\mu)\right]^2\\ &= \mathbb{E}(y_{ij}-\mu)^2+\mathbb{E}(\overline{y}_{i.}-\mu)^2-2\mathbb{E}\left[(y_{ij}-\mu)(\overline{y}_{i.}-\mu)\right]\\ &= \mathrm{Var}(y_{ij})+\mathrm{Var}(\overline{y}_{i.})-2\mathrm{Cov}(y_{ij},\overline{y}_{i.}) \ .\end{aligned}$$

We have calculated the variance of $\overline{y}_{i.}$ in (d). Using the covariance term obtained in the solution for (a),

$$\mathbb{E}\left(y_{ij}-\overline{y}_{i.}\right)^2 = \sigma_G^2+\sigma_E^2+\sigma_G^2+\frac{\sigma_E^2}{n_i}-2\sigma_G^2-2\frac{\sigma_E^2}{n_i}=\frac{n_i-1}{n_i}\sigma_E^2 \ . \tag{4}$$

Now substituting the expression for the summands obtained in (4) in (d),

$$\mathbb{E}[\mathrm{MS_W}] = \frac{1}{n-m}\sum_{i=1}^{m}(n_i-1)\sigma_E^2=\sigma_E^2 \ .$$

Similarly, for the $\mathrm{MS_B}$, from definition,

$$\mathrm{MS_B} = \frac{1}{m-1}\sum_{i=1}^{m}n_i\mathbb{E}\left(\overline{y}_{i.}-\overline{y}_{..}\right)^2 \ . \tag{5}$$

Analysing each summand in the RHS of (5),

$$\begin{aligned}\mathbb{E}\left(\overline{y}_{i.}-\overline{y}_{..}\right)^2 &= \mathbb{E}\left[(\overline{y}_{i.}-\mu)-(\overline{y}_{..}-\mu)\right]^2\\ &= \mathbb{E}(\overline{y}_{i.}-\mu)^2+\mathbb{E}(\overline{y}_{..}-\mu)^2-2\mathbb{E}\left[(\overline{y}_{i.}-\mu)(\overline{y}_{..}-\mu)\right]\\ &= \mathrm{Var}(\overline{y}_{i.})+\mathrm{Var}(\overline{y}_{..})-2\mathrm{Cov}(\overline{y}_{i.},\overline{y}_{..}) \ .\end{aligned}$$

We have already calculated the variance of $\overline{y}_{i.}$ in (d) and the variance of $\overline{y}_{..}$ in (c) and the covariance term can be obtained from (c). Therefore,

$$\begin{aligned}\mathbb{E}(\mathrm{MS_B}) &= \frac{1}{m-1}\sum_{i=1}^{m}n_i\mathbb{E}\left(\overline{y}_{i.}-\overline{y}_{..}\right)^2\\ &= \frac{1}{m-1}\left[\sum_{i=1}^{m}n_i\mathrm{Var}(\overline{y}_{i.})+n\mathrm{Var}(\overline{y}_{..})-2\sum_{i=1}^{m}n_i\mathrm{Cov}(\overline{y}_{i.},\overline{y}_{..})\right]\\ &= \frac{1}{m-1}\left[n\sigma_G^2+m\sigma_E^2+\sum_{i=1}^{m}\frac{n_i^2}{n}\sigma_G^2+\sigma_E^2-2\sum_{i=1}^{m}\frac{n_i^2}{n}\sigma_G^2-2\sigma_E^2\right] \qquad (6)\\ &= \frac{1}{m-1}\left[\left(n-\sum_{i=1}^{m}\frac{n_i^2}{n}\right)\sigma_G^2+(m-1)\sigma_E^2\right]\\ &= \frac{n-\sum_{i=1}^{m}\frac{n_i^2}{n}}{m-1}\sigma_G^2+\sigma_E^2 \ .\end{aligned}$$

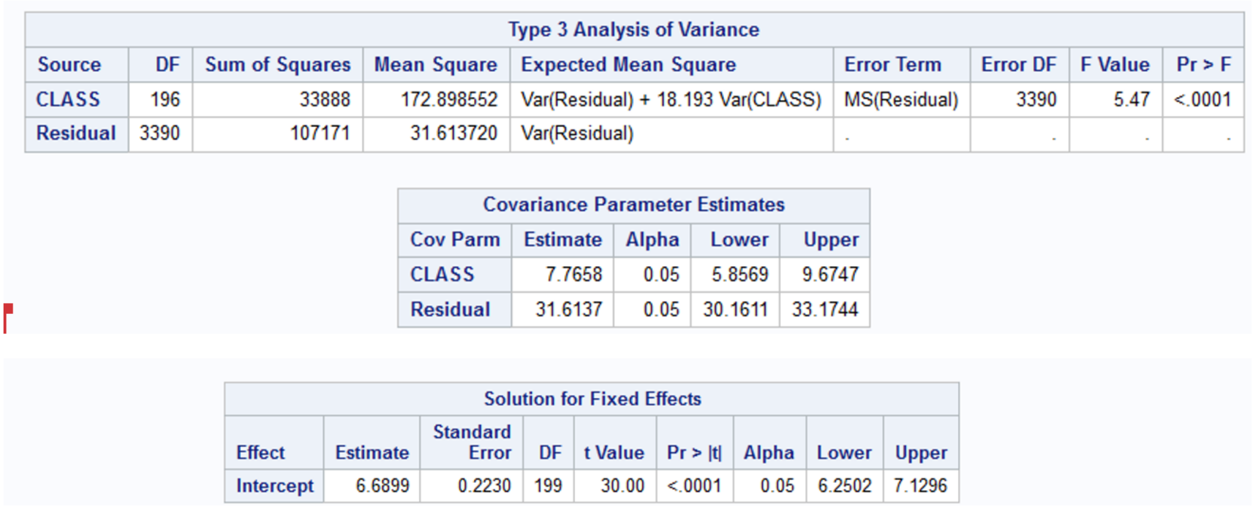
Solution A3.

(a) ANOVA results are shown in Figure 1.

(b) Confidence intervals for μ, σ_W^2 and σ_B^2 can be readily read of from these tables (see slides).

(c) The intraclass correlation and total variability can be calculated using the formulas on the slides. This yields an ICC of 19.7% [95%-CI of 16.1-24.1].

(d) The total variability of 39.4 [95%-CI of 37.1 - 41.8].



Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
CLASS	196	33888	172.898552	Var(Residual) + 18.193 Var(CLASS)	MS(Residual)	3390	5.47	<.0001
Residual	3390	107171	31.613720	Var(Residual)	.	.	.	.

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
CLASS	7.7658	0.05	5.8569	9.6747
Residual	31.6137	0.05	30.1611	33.1744

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	6.6899	0.2230	199	30.00	<.0001	0.05	6.2502	7.1296

Figure 1: Output for the one-way ANOVA on the difference in language test scores for all children with class as the random effect.

Solution A4.

(a) The results are tabulized below:

Class	$\bar{y}_{i.}$	Lowest	Class	$\bar{y}_{i.}$	Highest
11680	-1.47	-0.03	14780	14.36	13.17
1880	-0.63	0.44	11780	13.41	12.53
280	-2.29	1.01	982	14.81	12.16
1580	0.11	2.16	22880	13.09	12.09
21180	1.37	2.31	6580	13	12.01

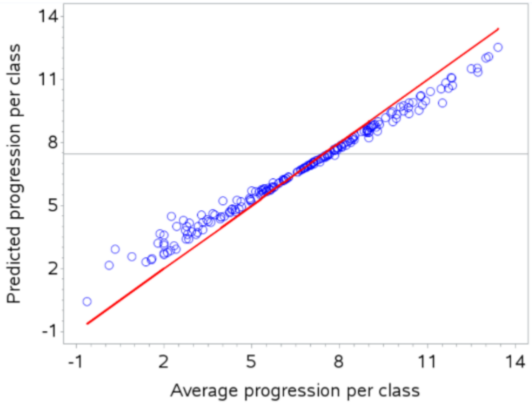
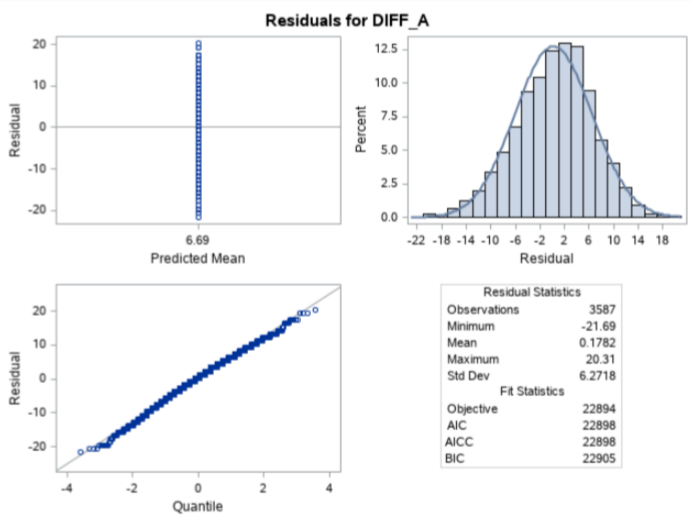


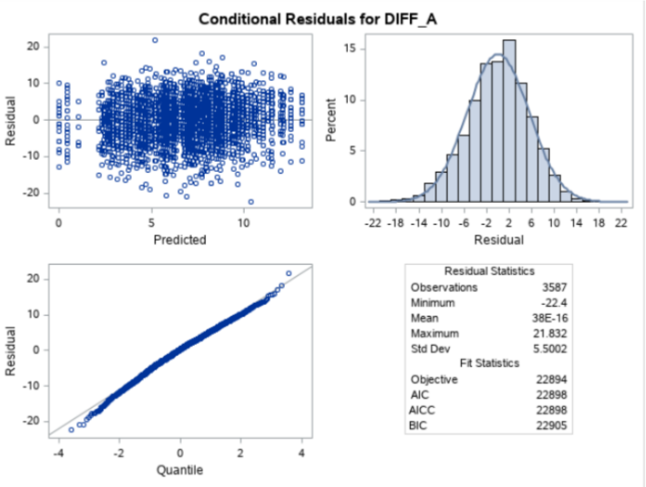
Figure 2: Preditions against Averages

(b) Note on the marginal there seems to be a bit of skewness to the right at the peak, and similarly in the conditional there is a peak slightly to the right of the peak of the normal dist. This can be used as reason to say it does not seem to follow the normal distribution. Overall the distribution does seem to be quite close to the normal distribution.



Residual Statistics	
Observations	3587
Minimum	-21.69
Mean	0.1782
Maximum	20.31
Std Dev	6.2718
Fit Statistics	
Objective	22894
AIC	22898
AICC	22898
BIC	22905

Figure 3: Marginal Residuals



Residual Statistics	
Observations	3587
Minimum	-22.4
Mean	38E-16
Maximum	21.832
Std Dev	5.5002
Fit Statistics	
Objective	22894
AIC	22898
AICC	22898
BIC	22905

Figure 4: Conditional Residuals

Solution A5.

(a) We rewrite the likelihood function given in slide 55 as

$$L(\mu, \sigma_G, \sigma_E) = \prod_{i=1}^{m} L_i(\mu, \sigma_G, \sigma_E),$$

where

$$\begin{aligned} L_i(\mu, \sigma_G, \sigma_E) &= \int_{-\infty}^{\infty} \frac{\phi\left(\frac{z}{\sigma_G}\right)}{\sigma_G} \prod_{j=1}^{n_i} \frac{1}{\sigma_E} \phi\left(\frac{y_{ij} - \mu - z}{\sigma_E}\right) dz \\ &= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}(n_i+1)} \sigma_G \sigma_E^{n_i}} \exp\left(-\frac{1}{2}\left(\left(\frac{z}{\sigma_G}\right)^2 + \sum_{j=1}^{n_i}\left(\frac{y_{ij} - \mu - z}{\sigma_E}\right)^2\right)\right) dz. \end{aligned}$$

We rewrite the sum of squares

$$\frac{1}{\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\mu-z)^2 = \frac{1}{\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2 + \frac{n_i}{\sigma_E^2}(\overline{y_{i.}}-\mu-z)^2,$$

so that we obtain

$$\begin{aligned}
L_i(\mu,\sigma_G,\sigma_E) &= \frac{1}{(2\pi)^{\frac{1}{2}(n_i+1)}\sigma_G\sigma_E^{n_i}}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2\right)\int\limits_{-\infty}^{\infty}\exp\left(-\frac{z^2}{2\sigma_G^2}-\frac{n_i}{2\sigma_E^2}(\overline{y_{i.}}-\mu-z)^2\right)dz \\
&= \frac{1}{(2\pi)^{\frac{1}{2}(n_i-1)}\sqrt{n_i}\sigma_E^{n_i-1}}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2\right)\int\limits_{-\infty}^{\infty}f(z;\sigma_G^2)f(\overline{y_{i.}}-\mu-z;\sigma_E^2/n_i)dz,
\end{aligned}$$

where $f(x;\sigma^2)=\exp(-x^2/2\sigma^2)/\sqrt{2\pi\sigma^2}$ is the probability density function of a normally distributed random variable with mean 0 and variance σ^2. It can be seen that the integral corresponds to the density of the sum of these two normally distributed random variables being equal to $\overline{y_{i.}}-\mu$. Since the sum of two normally distributed random variables is again normally distributed with variance equal to the sum of the variances, the integral evaluates to $f(\overline{y_{i.}}-\mu;\sigma_G^2+\sigma_E^2/n_i)$. Therefore, we have

$$\begin{aligned}
L_i(\mu,\sigma_G,\sigma_E) &= \frac{1}{(2\pi)^{\frac{1}{2}(n_i-1)}\sqrt{n_i}\sigma_E^{n_i-1}}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2\right)\frac{1}{\sqrt{2\pi\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}}\exp\left(-\frac{(\overline{y_{i.}}-\mu)^2}{2\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\right) \\
&= \frac{1}{(2\pi)^{n_i/2}\sigma_E^{n_i-1}\sqrt{n_i\sigma_G^2+\sigma_E^2}}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2-\frac{(\overline{y_{i.}}-\mu)^2}{2\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\right).
\end{aligned}$$

To find the μ that maximizes the total likelihood $L(\mu,\sigma_G,\sigma_E)$, we take the derivative of the log-likelihood

$$\frac{d}{d\mu}\log L(\mu,\sigma_G,\sigma_E) = \sum_{i=1}^{m}\frac{d}{d\mu}\log L_i(\mu,\sigma_G,\sigma_E) = \sum_{i=1}^{m}\frac{\frac{d}{d\mu}L_i(\mu,\sigma_G,\sigma_E)}{L_i(\mu,\sigma_G,\sigma_E)}. \tag{7}$$

Differentiating L_i w.r.t. μ, yields

$$\begin{aligned}
\frac{d}{d\mu}L_i(\mu,\sigma_G,\sigma_E) &= \frac{1}{(2\pi)^{n_i/2}\sqrt{n_i\sigma_G^2+\sigma_E^2}\sigma_E^{n_i-1}}\frac{\overline{y_{i.}}-\mu}{\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2-\frac{(\overline{y_{i.}}-\mu)^2}{2\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\right) \\
&= \frac{\overline{y_{i.}}-\mu}{\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}L_i(\mu,\sigma_G,\sigma_E).
\end{aligned}$$

We substitute this into (7) to find the zero of this derivative

$$\sum_{i=1}^{m}\frac{\frac{d}{d\mu}L_i(\mu,\sigma_G,\sigma_E)}{L_i(\mu,\sigma_G,\sigma_E)} = \sum_{i=1}^{m}\frac{\overline{y_{i.}}-\mu}{\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)} = 0.$$

Solving this, we obtain

$$\mu = \frac{\sum_{i=1}^{m}\overline{y_{i.}}/\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}{\sum_{i=1}^{m}1/\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)} = \frac{\sum_{i=1}^{m}\overline{y_{i.}}n_i\sigma_E^2/\left(n_i\sigma_G^2+\sigma_E^2\right)}{\sum_{i=1}^{m}n_i\sigma_E^2/\left(n_i\sigma_G^2+\sigma_E^2\right)},$$

as required. We now take the derivative w.r.t. σ_G

$$\frac{d}{d\sigma_G}L_i(\mu,\sigma_G,\sigma_E) = \frac{1}{(2\pi)^{n_i/2}\sigma_E^{n_i-1}}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2\right)\frac{d}{d\sigma_G}\frac{1}{\sqrt{n_i\sigma_G^2+\sigma_E^2}}\exp\left(-\frac{(\overline{y_{i.}}-\mu)^2}{2\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\right).$$

The derivative of the first factor is

$$\frac{-n_i\sigma_G}{(n_i\sigma_G^2+\sigma_E^2)^{3/2}},$$

while the derivative of the second factor is

$$\frac{\sigma_G(\overline{y_{i.}}-\mu)^2}{(\sigma_G^2+\sigma_E^2/n_i)^2}\exp\left(-\frac{(\overline{y_{i.}}-\mu)^2}{2\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\right).$$

Bringing these together, we get

$$\frac{d}{d\sigma_G}L_i(\mu,\sigma_G,\sigma_E)=\frac{-n_i\sigma_G}{n_i\sigma_G^2+\sigma_E^2}L_i(\mu,\sigma_G,\sigma_E)+\frac{\sigma_G(\overline{y_{i.}}-\mu)^2}{(\sigma_G^2+\sigma_E^2/n_i)^2}L_i(\mu,\sigma_G,\sigma_E).$$

So that

$$\frac{d}{d\sigma_G}L(\mu,\sigma_G,\sigma_E)=\sum_{i=1}^m\frac{\frac{d}{d\sigma_G}L_i(\mu,\sigma_G,\sigma_E)}{L_i(\mu,\sigma_G,\sigma_E)}=\sum_{i=1}^m\frac{-n_i\sigma_G}{n_i\sigma_G^2+\sigma_E^2}+\frac{\sigma_G(\overline{y_{i.}}-\mu)^2}{(\sigma_G^2+\sigma_E^2/n_i)^2}.$$

Setting this to zero, dividing by σ_G, and moving the negative terms to the other side gives

$$\sum_{i=1}^m\frac{n_i^2(\overline{y_{i.}}-\mu)^2}{(n_i\sigma_G^2+\sigma_E^2)^2}=\sum_{i=1}^m\frac{n_i}{n_i\sigma_G^2+\sigma_E^2},$$

as required. The last equation is obtained by taking the derivative w.r.t. σ_E.

$$\frac{d}{d\sigma_E}L_i(\mu,\sigma_G,\sigma_E)=(2\pi)^{-n_i/2}\frac{d}{d\sigma_E}\frac{1}{\sigma_E^{n_i-1}}\frac{1}{\sqrt{n_i\sigma_G^2+\sigma_E^2}}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2\right)\exp\left(-\frac{(\overline{y_{i.}}-\mu)^2}{2\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\right).$$

We separately take the derivatives of these four factors:

$$\begin{aligned}
\frac{d}{d\sigma_E}\frac{1}{\sigma_E^{n_i-1}}&=-\frac{n_i-1}{\sigma_E^{n_i}}=-\frac{n_i-1}{\sigma_E}\frac{1}{\sigma_E^{n_i-1}},\\
\frac{d}{d\sigma_E}\frac{1}{\sqrt{n_i\sigma_G^2+\sigma_E^2}}&=-\frac{\sigma_E}{(n_i\sigma_G^2+\sigma_E)^{3/2}}=-\frac{\sigma_E}{n_i\sigma_G^2+\sigma_E}\frac{1}{\sqrt{n_i\sigma_G^2+\sigma_E^2}},\\
\frac{d}{d\sigma_E}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2\right)&=\frac{\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2}{\sigma_E^3}\exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2\right),\\
\frac{d}{d\sigma_E}\exp\left(-\frac{(\overline{y_{i.}}-\mu)^2}{2\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\right)&=\frac{\sigma_E n_i(\overline{y_{i.}}-\mu)^2}{(n_i\sigma_G^2+\sigma_E^2)^2}\exp\left(-\frac{(\overline{y_{i.}}-\mu)^2}{2\left(\sigma_G^2+\frac{\sigma_E^2}{n_i}\right)}\right).
\end{aligned}$$

The derivative of the likelihood then becomes

$$\begin{aligned}
\frac{d}{d\sigma_E}L(\mu,\sigma_G,\sigma_E)&=\sum_{i=1}^m\frac{\frac{d}{d\sigma_E}L_i(\mu,\sigma_G,\sigma_E)}{L_i(\mu,\sigma_G,\sigma_E)}\\
&=\sum_{i=1}^m-\frac{n_i-1}{\sigma_E}-\frac{\sigma_E}{n_i\sigma_G^2+\sigma_E}+\frac{\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2}{\sigma_E^3}+\frac{\sigma_E n_i(\overline{y_{i.}}-\mu)^2}{(n_i\sigma_G^2+\sigma_E^2)^2}.
\end{aligned}$$

We set this to zero, divide by σ_E, and move the negative terms to the other side to obtain

$$\sum_{i=1}^m\sum_{j=1}^{n_i}\frac{(y_{ij}-\overline{y_{i.}})^2}{\sigma_E^4}+\sum_{i=1}^m\frac{n_i(\overline{y_{i.}}-\mu)^2}{(n_i\sigma_G^2+\sigma_E^2)^2}=\sum_{i=1}^m\frac{n_i-1}{\sigma_E^2}+\sum_{i=1}^m\frac{1}{n_i\sigma_G^2+\sigma_E}.$$

The desired result is then obtained by noting that the double sum equals SS_w while $\sum_{i=1}^m n_i-1=n-m$.

(b) The ML solutions solve the likelihood equations. We assume the data is balanced, so $n_i = N/m = n_0$. Start with the likelihood equations (slide 55). We have

$$\begin{aligned}
\mu &= \frac{\sum_{i=1}^m \sigma_E^2 \bar{y}_{i.} / \left[n_i \sigma_G^2 + \sigma_E^2\right]}{\sum_{i=1}^m \sigma_E^2 / \left[n_i \sigma_G^2 + \sigma_E^2\right]} \\
\sum_{i=1}^m \left[\frac{n_i^2 \left(\bar{y}_{i.} - \mu\right)^2}{\left[n_i \sigma_G^2 + \sigma_E^2\right]^2}\right] &= \sum_{i=1}^m \left[\frac{n_i}{n_i \sigma_G^2 + \sigma_E^2}\right] \\
\sum_{i=1}^m \left[\frac{n_i \left(\bar{y}_i - \mu\right)^2}{\left[n_i \sigma_G^2 + \sigma_E^2\right]^2}\right] + \frac{SS_W}{\sigma_E^4} &= \sum_{i=1}^m \left[\frac{1}{n_i \sigma_G^2 + \sigma_E^2}\right] + \frac{n-m}{\sigma_E^2}
\end{aligned}$$

Since $n_i = n_0$ is fixed, in the first equation $\frac{\sigma_E^2}{(n_0\sigma_G^2 + \sigma_E^2)}$ is constant. Hence,

$$\hat{\mu} = \frac{\sigma_E^2 / [n_0 \sigma_G^2 + \sigma_E^2]}{\sigma_E^2 / [n_0 \sigma_G^2 + \sigma_E^2]} \frac{\sum_{i=1}^m \bar{y}_{i.}}{\sum_{i=1}^m 1} = \frac{\sum_{i=1}^m y_{i.}}{m} = \bar{y}_{..}$$

In the second equation, $n_0\sigma_G^2 + \sigma_E^2$ is constant. Multiply both sides by it, to find

$$\begin{aligned}
&\sum_{i=1}^m n_0^2 (\bar{y}_i - \mu)^2 = (n_0 \sigma_G^2 + \sigma_E^2) \sum_{i=1}^m n_0 \\
\Longleftrightarrow\ & n_0 \sum_{i=1}^m n_0 (\bar{y}_i - \mu)^2 = n_0 (n_0 \sigma_G^2 + \sigma_E^2) \sum_{i=1}^m 1 \\
\Longleftrightarrow\ & \sum_{i=1}^m n_0 (\bar{y}_i - \mu)^2 = (n_0 \sigma_G^2 + \sigma_E^2) \sum_{i=1}^m 1 \\
\Longleftrightarrow\ & \sum_{i=1}^m n_0 (\bar{y}_i - \mu)^2 = n_0 m \sigma_G^2 + m \sigma_E^2 \\
\Longleftrightarrow\ & SS_b = N\sigma_G^2 + m\sigma_e^2,
\end{aligned}$$

where the final step uses that $y_{..} = \mu$ by the argument above. This has two unknowns, so we cannot derive the likelihood estimator just yet.

Turning to the third equation, we can again take denominators out since n_0 is fixed. This yields

$$\begin{aligned}
& \frac{1}{\left[n_0 \sigma_G^2 + \sigma_E^2\right]^2} \sum_{i=1}^m \left[n_0 \left(\bar{y}_i - \mu\right)^2\right] + \frac{SS_W}{\sigma_E^4} = \frac{m}{n_0 \sigma_G^2 + \sigma_E^2} + \frac{N-m}{\sigma_E^2} \\
\Longleftrightarrow\ & \frac{1}{\left[n_0 \sigma_G^2 + \sigma_E^2\right]^2} SS_B + \frac{SS_W}{\sigma_E^4} = \frac{m}{n_0 \sigma_G^2 + \sigma_E^2} + \frac{N-m}{\sigma_E^2} \\
\Longleftrightarrow\ & \frac{SS_B - m(n_0 \sigma_G^2 + \sigma_E^2)}{\left[n_0 \sigma_G^2 + \sigma_E^2\right]^2} = \frac{(N-m)\sigma_E^2 - SS_W}{\sigma_E^4}.
\end{aligned}$$

Using that $SS_b = N\sigma_G^2 + m\sigma_e^2$ (see above), the left hand side of this equation reduces to 0. Hence,

$$\frac{(N-m)\sigma_E^2 - SS_W}{\sigma_E^4} = 0 \iff (N-m)\sigma_E^2 = SS_W \iff \hat{\sigma}_E^2 = \frac{SS_W}{N-m} = MS_W$$

We can use this result together with $SS_b = N\sigma_G^2 + m\sigma_E^2$ to find that

$$\hat{\sigma}_G^2 = \frac{1}{n_0}\left[\frac{m-1}{m} MS_b - MS_w\right]$$

(c) The ML *estimators* solve the likelihood equations under the restriction that $\hat{\sigma}_G \geq 0$. Note that the likelihood *solutions* do not solve it under this restriction, as $\hat{\sigma}_G^2$ becomes negative whenever $\frac{m-1}{m}MS_b - MS_w < 0$.

In case $\frac{m-1}{m}MS_b - MS_w < 0$, we truncate $\hat{\sigma}_G^2$ to the boundary point of 0. Hence,

$$\hat{\sigma}_G^2 = \max\left\{0, \frac{1}{n_0}\left[\frac{m-1}{m}MS_b - MS_w\right]\right\}$$

In case $\frac{m-1}{m}MS_b - MS_w < 0$, thus $\sigma_G = 0$. Plugging this in in the third likelihood equation followed by multiplication of both sides with σ_E^4 yields

$$\sum_{i=1}^{m} n_i(\bar{y}_{i.} - \mu)^2 + SS_w = \sigma_E^2\,[m + (N-m)]].$$

The MLE solution of μ is not affected by the truncation. Hence, $\hat{\mu} = y_{..}$. Plugging this in above, we find that

$$SS_B + SS_W = N\sigma_E^2.$$

Hence, indeed if $\frac{m-1}{m}MS_b - MS_w < 0$, we have $\hat{\sigma}_E^2 = \frac{1}{mn_0}(SS_b + SS_w)$ if we truncate.

(d) The likelihood and log-likelihood function for balanced data is:

$$\begin{aligned}
L &= \prod_{i=1}^{m}\left(\frac{1}{\sigma_E\sqrt{2\pi}}\right)^{n_0}\frac{1}{\sigma_G}\frac{\sigma_E\sigma_G}{\sqrt{n_0\sigma_G^2+\sigma_E^2}}\exp\left(\frac{n_0^2\sigma_G^2(\overline{y_{i.}}-\mu)^2}{2\sigma_E^2(n_0\sigma_G^2+\sigma_E^2)} - \frac{1}{2\sigma_E^2}\sum_{j=1}^{n_0}(y_{ij}-\mu)^2\right)\\
&= \sigma_E^{-m(n_0-1)}(2\pi)^{-\frac{mn_0}{2}}(n_0\sigma^2+\sigma_E^2)^{-\frac{m}{2}}\exp\left(\sum_{i=1}^{m}\frac{n_0^2\sigma_G^2(\overline{y_{i.}}-\mu)^2}{2\sigma_E^2(n_0\sigma_G^2+\sigma_E^2)} - \frac{1}{2\sigma_E^2}\sum_{i=1}^{m}\sum_{j=1}^{n_0}(y_{ij}-\mu)^2\right)\\
\log(L) &= -\frac{m(n_0-1)}{2}\log(\sigma_E^2) - \frac{mn_0}{2}\log(2\pi) - \frac{m}{2}\log(n_0\sigma_G^2+\sigma_E^2) + \sum_{i=1}^{m}\frac{n_0^2\sigma_G^2(\overline{y_{i.}}-\mu)^2}{2\sigma_E^2(n_0\sigma_G^2+\sigma_E^2)} - \frac{1}{2\sigma_E^2}\sum_{i=1}^{m}\sum_{j=1}^{n_0}(y_{ij}-\mu)^2
\end{aligned}$$

Next we determine the first derivatives: (Note make sure to derive over σ_G^2, not σ_G)

$$\begin{aligned}
\frac{\delta}{\delta\mu}\log(L) &= -\sum_{i=1}^{m}\frac{n_0^2\sigma_G^2(\overline{y_{i.}}-\mu)}{\sigma_E^2(n_0\sigma_E^2+\sigma_G^2)} + \frac{1}{\sigma_E^2}\sum_{i=1}^{m}\sum_{j=1}^{n_0}(y_{ij}-\mu) = -\frac{mn_0^2\sigma_G^2(\overline{y_{..}}-\mu)}{\sigma_E^2(n_0\sigma_E^2+\sigma_G^2)} + \frac{mn_0}{\sigma_E^2}(\overline{y_{..}}-\mu)\\
&= \frac{mn_0}{n_0\sigma_G^2+\sigma_E^2}(\overline{y_{..}}-\mu)\\
\frac{\delta}{\delta\sigma_G^2}\log(L) &= -\frac{mn_0}{2(n_0\sigma_G^2+\sigma_E^2)} + \sum_{i=1}^{m}\frac{n_0^2(\overline{y_{i.}}-\mu)^2}{2(n_0\sigma_G^2+\sigma_E^2)^2}\\
\frac{\delta}{\delta\sigma_E^2}\log(L) &= -\frac{m(n_0-1)}{2\sigma_E^2} - \frac{m}{2(n_0\sigma_G^2+\sigma_E^2)} - \sum_{i=1}^{m}\frac{n_0^2\sigma_G^2(n_0\sigma_G^2+2\sigma_E^2)}{2\sigma_E^4(n_0\sigma_G^2+\sigma_E^2)^2}(\overline{y_{i.}}-\mu)^2 + \frac{1}{2\sigma_E^4}\sum_{i=1}^{m}\sum_{j=1}^{n_0}(y_{ij}-\mu)^2
\end{aligned}$$

Then the second derivatives are:

$$\begin{aligned}
\frac{\delta^2}{\delta\mu^2}\log(L) &= \frac{-mn_0}{n_0\sigma_G^2+\sigma_E^2}\\
\frac{\delta^2}{\delta\mu\sigma_G^2}\log(L) &= \frac{-mn_0^2}{(n_0\sigma_G^2+\sigma_E^2)^2}(\overline{y_{..}}-\mu)\\
\frac{\delta^2}{\delta\mu\sigma_E^2}\log(L) &= \frac{-mn_0}{(n_0\sigma_G^2+\sigma_E^2)^2}(\overline{y_{..}}-\mu)\\
\frac{\delta^2}{\delta\sigma_G^4}\log(L) &= \frac{mn_0^2}{2(n_0\sigma_G^2+\sigma_E^2)^2}-\sum_{i=1}^{m}\frac{n_0^3(\overline{y_{i.}}-\mu)^2}{(n_0\sigma_G^2+\sigma_E^2)^3}\\
\frac{\delta^2}{\delta\sigma_G^2\sigma_E^2}\log(L) &= \frac{mn_0}{2(n_0\sigma_G^2+\sigma_E^2)^2}-\sum_{i=1}^{m}\frac{n_0^2(\overline{y_{i.}}-\mu)^2}{(n_0\sigma_G^2+\sigma_E^2)^3}\\
\frac{\delta^2}{\delta\sigma_E^4}\log(L) &= \frac{m(n_0-1)}{2\sigma_E^4}+\frac{m}{2(n_0\sigma_G^2+\sigma_E^2)^2}+\sum_{i=1}^{m}\left(\frac{1}{\sigma_E^6}-\frac{1}{({}_0\sigma_G^2+\sigma_E^2)^3}\right)n_0(\overline{y_{i.}}-\mu)^2\\
&\quad-\frac{1}{\sigma_E^6}\sum_{i=1}^{m}\sum_{j=1}^{n_0}(y_{ij}-\mu)^2
\end{aligned}$$

Now we know the following:

$$\begin{aligned}
&\mathbb{E}[\overline{y_{..}}-\mu]=0\\
&\mathbb{E}[(\overline{y_{i.}}-\mu)^2]=\mathrm{Var}(\overline{y_{i.}})=\sigma_G^2+\sigma_E^2/n_0\\
&\mathbb{E}[\sum_{i=1}^{m}\sum_{j=1}^{n_0}(y_{ij}-\mu)^2]=mn_0\mathrm{Var}(y_{ij})=mn_0(\sigma_G^2+\sigma_E^2)
\end{aligned}$$

Filling this in gives the Fisher matrix F_z, where $n_0\sigma^2+\sigma_E^2=\lambda$:

$$F_z=\begin{pmatrix}\frac{mn_0}{\lambda} & 0 & 0\\ 0 & \frac{mn_0^2}{2\lambda^2} & \frac{mn_0}{2\lambda^2}\\ 0 & \frac{mn_0}{2\lambda^2} & \frac{m((n_0-1)\lambda^2+\sigma_E^4)}{2\sigma_E^4\lambda^2}\end{pmatrix}$$

(e)

$$F_z^{-1}=\begin{pmatrix}\frac{\lambda}{mn_0} & 0 & 0\\ 0 & \frac{2(\lambda^2(n_0-1)+\sigma_E^4)}{mn_0^2(n_0-1)} & \frac{-2\sigma_E^4}{mn_0(n_0-1)}\\ 0 & \frac{-2\sigma_E^4}{mn_0(n_0-1)} & \frac{-2\sigma_E^4}{m(n_0-1)}\end{pmatrix}=\begin{pmatrix}\frac{\lambda}{mn_0} & 0 & 0\\ 0 & \frac{2\sigma_E^4}{mn_0^2}\left(\frac{1}{n_0-1}+\frac{\lambda^2}{\sigma_E^4}\right) & \frac{-2\sigma_E^4}{mn_0(n_0-1)}\\ 0 & \frac{-2\sigma_E^4}{mn_0(n_0-1)} & \frac{-2\sigma_E^4}{m(n_0-1)}\end{pmatrix}$$

(f)-(g) Estimating the model in SAS results in the following:

Variable	Estimate	LCL	UCL
μ	6.6893	6.2475	7.1311
σ_G^2	7.8554	6.2194	10.2375
σ_E^2	31.6176	30.1650	33.1782

(h)-(i) To answer the questions about the confidence limits for the ICC and total variability we refer to slides 74-75. For the ICC we have that:

$$\begin{aligned}
&\sigma_G^2=7.8554,\sigma_E^2=31.6176,\\
&\tau_{GE}=\tau_{EG}=-0.03695,\tau_{GG}=0.9896,\tau_{EE}=0.5896.
\end{aligned}$$

For the ICC we can immediately use this (ignoring the covariance) resulting in an estimate of 0.19901 and CL: $[0.15959, 0.23978]$.

For the total variability we have to also consider the covariance and see that: $\hat{\tau}^2 = 0.5896 + 0.9896 + 2 \cdot (-0.03695)$. And find an estimate for the total variability of: 39.473 with confidence limits: $[37.1745, 41.9925]$.

Solution A6.

(a) In A5, it has been shown that the likelihood function is given by

$$L(\mu, \sigma_G, \sigma_E) = \prod_{i=1}^{m} \frac{1}{(2\pi)^{n_i/2}\sigma_E^{n_i-1}\sqrt{n_i\sigma_G^2 + \sigma_E^2}} \exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i.}})^2 - \frac{(\overline{y_{i.}} - \mu)^2}{2\left(\sigma_G^2 + \frac{\sigma_E^2}{n_i}\right)}\right).$$

To obtain the REML likelihood function, we integrate over μ. For this, we first rewrite the factor that contains μ:

$$\exp\left(-\sum_{i=1}^{m} \frac{(\overline{y_{i.}} - \mu)^2}{2\left(\sigma_G^2 + \frac{\sigma_E^2}{n_i}\right)}\right).$$

We rewrite the exponent as

$$\begin{aligned}
-\sum_{i=1}^{m} \frac{(\overline{y_{i.}} - \mu)^2}{2\left(\sigma_G^2 + \frac{\sigma_E^2}{n_i}\right)} &= -\frac{1}{2}\left(\mu^2 \sum_{i=1}^{m} \frac{n_i}{n_i\sigma_G^2 + \sigma_E^2} + \sum_{i=1}^{m} \overline{y_{i.}}^2 \frac{n_i}{n_i\sigma_G^2 + \sigma_E^2} - 2\mu \sum_{i=1}^{m} \overline{y_{i.}} \frac{n_i}{n_i\sigma_G^2 + \sigma_E^2}\right) \\
&= -\frac{1}{2}\left(\mu^2 \sum_{i=1}^{m} \alpha_i + \sum_{i=1}^{m} \alpha_i \overline{y_{i.}}^2 - 2\mu \sum_{i=1}^{m} \alpha_i \overline{y_{i.}}\right),
\end{aligned}$$

where $\alpha_i = n_i/(n_i\sigma_G^2 + \sigma_E^2)$. This, in turn, can be rewritten to

$$-\frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i\right)\left(\mu - \frac{\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}}{\sum_{i=1}^{m} \alpha_i}\right)^2 + \frac{\left(\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}\right)^2}{2\sum_{i=1}^{m} \alpha_i} - \frac{1}{2}\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}^2.$$

We then take the integral over the part of the exponential that contains μ (as all other terms are constants that can be taken outside of the integral):

$$\begin{aligned}
\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i\right)\left(\mu - \frac{\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}}{\sum_{i=1}^{m} \alpha_i}\right)^2\right) d\mu &= \int_{-\infty}^{\infty} \sqrt{2\pi}\phi\left(\sqrt{\sum_{i=1}^{m} \alpha_i}\left(\mu - \frac{\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}}{\sum_{i=1}^{m} \alpha_i}\right)\right) d\mu \\
&= \sqrt{\frac{2\pi}{\sum_{i=1}^{m} \alpha_i}}.
\end{aligned}$$

Now, when putting back all other terms, we obtain

$$\begin{aligned}
&\int_{-\infty}^{\infty} L(\mu, \sigma_G, \sigma_E) d\mu \\
=&\sqrt{\frac{2\pi}{\sum_{i=1}^{m} \alpha_i}} \exp\left(\frac{\left(\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}\right)^2}{2\sum_{i=1}^{m} \alpha_i} - \frac{1}{2}\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}^2\right) \prod_{i=1}^{m} \frac{1}{(2\pi)^{n_i/2}\sigma_E^{n_i-1}\sqrt{n_i\sigma_G^2 + \sigma_E^2}} \exp\left(-\frac{1}{2\sigma_E^2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i.}})^2\right) \\
=&\frac{\exp\left(-\frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}^2 - \frac{\left(\sum_{i=1}^{m} \alpha_i \overline{y_{i.}}\right)^2}{\sum_{i=1}^{m} \alpha_i} + \frac{1}{\sigma_E^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i.}})^2\right)\right)}{(2\pi)^{(n-1)/2}\sigma_E^{n-m}\sqrt{\sum_{i=1}^{m} \alpha_i}\prod_{i=1}^{m}\sqrt{\alpha_i/n_i}}.
\end{aligned}$$

The corresponding log-likelihood is given by

$$\begin{aligned}
&\log\left(\int_{-\infty}^{\infty} L(\mu,\sigma_G,\sigma_E)d\mu\right)\\
&= -\frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i\overline{y_{i.}}^2 - \frac{\left(\sum_{i=1}^{m}\alpha_i\overline{y_{i.}}\right)^2}{\sum_{i=1}^{m}\alpha_i} + \frac{SS_W}{\sigma_E^2}\right)\\
&\quad - \frac{n-1}{2}\log 2\pi - \frac{n-m}{2}\log\sigma_E^2 - \frac{1}{2}\log\sum_{i=1}^{m}\alpha_i + \frac{1}{2}\sum_{i=1}^{m}\log(\alpha_i) - \frac{1}{2}\sum_{i=1}^{m}\log n_i.
\end{aligned}$$

(b) Using:

$$\begin{aligned}
\frac{\delta}{\delta\sigma_G^2}\alpha_i &= \frac{\delta}{\delta\sigma_G^2}\frac{n_i}{n_i\sigma_G^2+\sigma_E^2} = -\alpha_i^2\\
\frac{\delta}{\delta\sigma_E^2}\alpha_i &= -\frac{\alpha_i^2}{n_i}
\end{aligned}$$

The first derivatives of the log-likelihood are:

$$\begin{aligned}
\frac{\delta}{\delta\sigma_G^2}\log(\text{REML}) &= \frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i^2\overline{y_{i.}}^2 + \left(\frac{\sum_{i=1}^{m}\alpha_i\overline{y_{i.}}}{\sum_{i=1}^{m}\alpha_i}\right)^2\sum_{i=1}^{m}\alpha_i^2 - 2\frac{\sum_{i=1}^{m}\alpha_i\overline{y_{i.}}}{\sum_{i=1}^{m}\alpha_i}\sum_{i=1}^{m}\alpha_i^2\overline{y_{i.}}\right)\\
&\quad + \frac{1}{2}\frac{\sum_{i=1}^{m}\alpha_i^2}{\sum_{i=1}^{m}\alpha_i} - \frac{1}{2}\sum_{i=1}^{m}\alpha_i\\
\frac{\delta}{\delta\sigma_E^2}\log(\text{REML}) &= \frac{1}{2}\left(\sum_{i=1}^{m}\frac{\alpha_i^2}{n_i}\overline{y_{i.}}^2 + \left(\frac{\sum_{i=1}^{m}\alpha_i\overline{y_{i.}}}{\sum_{i=1}^{m}\alpha_i}\right)^2\sum_{i=1}^{m}\frac{\alpha_i^2}{n_i} - 2\frac{\sum_{i=1}^{m}\alpha_i\overline{y_{i.}}}{\sum_{i=1}^{m}\alpha_i}\sum_{i=1}^{m}\frac{\alpha_i^2}{n_i}\overline{y_{i.}} + \frac{\text{SS}_W}{\sigma_E^4}\right)\\
&\quad - \frac{n-m}{2\sigma_E^2} + \frac{1}{2}\frac{\sum_{i=1}^{m}\frac{\alpha_i^2}{n_i}}{\sum_{i=1}^{m}\alpha_i} - \frac{1}{2}\sum_{i=1}^{m}\frac{\alpha_i}{n_i}
\end{aligned}$$

The likelihood equations are:

$$\frac{\delta}{\delta\sigma_G^2}\log(\text{REML}) = 0 \qquad\qquad \frac{\delta}{\delta\sigma_E^2}\log(\text{REML}) = 0$$

(c) Now $\alpha_i = \alpha$ giving the following likelihood equations:

$$\begin{aligned}
\frac{2}{\alpha}\frac{\delta}{\delta\sigma_G^2}\log(\text{REML}) &= \alpha\sum_{i=1}^{m}\left(\overline{y_{i.}}^2 - \overline{y_{..}}^2\right) - (m-1) = \alpha\sum_{i=1}^{m}\left(\overline{y_{i.}} - \overline{y_{..}}\right)^2 - (m-1) = \alpha\frac{\text{SS}_B}{n_0} - (m-1) = 0\\
\alpha &= \frac{n_0}{\text{MS}_B}\\
2\frac{\delta}{\delta\sigma_E^2}\log(\text{REML}) &= \frac{\alpha}{n_0}\left(\frac{\alpha\text{SS}_B}{n_0} - (m-1)\right) + \frac{\text{SS}_W}{\sigma_E^4} - \frac{n-m}{\sigma_E^2} = 0
\end{aligned}$$

Solving this gives:

$$\hat{\sigma_E}^2 = \text{MS}_W \qquad\qquad \hat{\sigma_G}^2 = \frac{\text{MS}_B - \text{MS}_W}{n_0}$$

Maximizing $\log(L(\mu,\hat{\sigma_G},\hat{\sigma_E}))$ w.r.t. μ gives $\hat{\mu} = \overline{y_{..}}$.

(d) Similar to A5, $\hat{\sigma}_G^2 \geq 0$:

$$\hat{\sigma}_G^2 = \max\{0, \frac{\text{MS}_B - \text{MS}_W}{n_0}\}$$

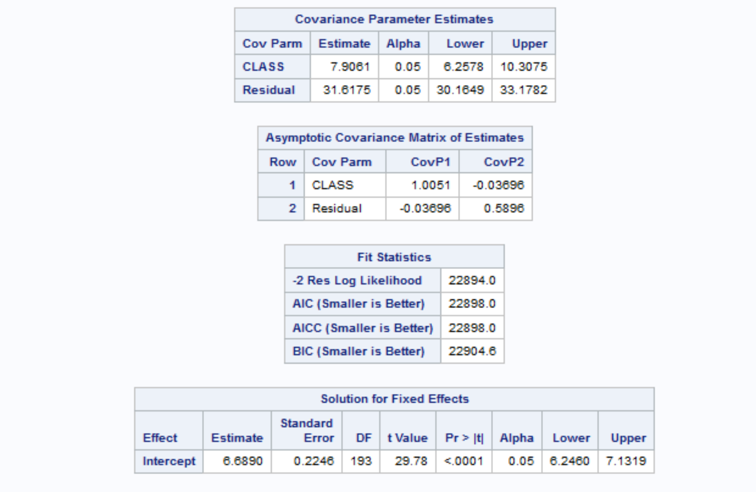
Filling in $\hat{\sigma}_G^2 = 0$, then $\alpha = \frac{n_0}{\sigma_E^2}$. Filling this in in the second likelihood equation (note $n = mn_0$):

$$2\sigma_E^2\frac{\delta}{\delta\sigma_E^2}\log(\text{REML}) = \frac{\text{SS}_B + \text{SS}_W}{\sigma_E^2} - (n-1) = 0$$

This gives the following REML solutions:

$$\hat{\mu} = \overline{y_{..}}$$
$$\hat{\sigma}_G^2 = \max\{0, \frac{\text{MS}_B - \text{MS}_W}{n_0}\}$$
$$\hat{\sigma}_G^2 = \begin{cases} \frac{\text{SS}_B + \text{SS}_W}{mn_0 - 1} & \text{if } \hat{\sigma}_G^2 = 0 \\ \text{MS}_W & \text{else} \end{cases}$$

(e)-(h) For the data analytical questions we use the same methods as in A5. The output can be found below:



Covariance Parameter Estimates				
Cov Parm	Estimate	Alpha	Lower	Upper
CLASS	7.9061	0.05	6.2578	10.3075
Residual	31.6175	0.05	30.1649	33.1782

Asymptotic Covariance Matrix of Estimates			
Row	Cov Parm	CovP1	CovP2
1	CLASS	1.0051	-0.03696
2	Residual	-0.03696	0.5896

Fit Statistics	
-2 Res Log Likelihood	22894.0
AIC (Smaller is Better)	22898.0
AICC (Smaller is Better)	22898.0
BIC (Smaller is Better)	22904.6

Solution for Fixed Effects								
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	6.6890	0.2246	193	29.78	<.0001	0.05	6.2460	7.1319

The results now are:

Variable	Estimate	LCL	UCL
μ	6.6890	6.2460,	7.1319
σ_G^2	7.9061	6.2578	10.3075
σ_E^2	31.6175	30.1649	33.1782
ICC	0.20003	0.16040	0.24100
Total Variability	39.5236	37.2137	42.0565

Solution A7.

(a) Estimating the model using TYPE1 we obtain the results in Table 1. For TYPE3 we get the results given in Table 2.

Effect	DF	SS	MS
COMBI	1	423.76	423.76
CLASS(COMBI)	185	30608	165.44
RES	3227	102487	31.76

Table 1: Results for TYPE1 estimation of the two-way nested mixed effects model.

Effect	DF	SS	MS
COMBI	1	767.66	767.66
CLASS(COMBI)	185	30608	165.44
RES	3227	102487	31.76

Table 2: Results for TYPE3 estimation of the two-way nested mixed effects model.

(b) The parameter estimates and the related confidence intervals coincide for TYPE1 and TYPE3 estimation. These are given in Table 3.

Variable	Estimate	LCL	UCL
μ	6.03	5.28	6.79
$\alpha_{COMBI=1}$	1.06	0.13	2.00
σ_G^2	7.32	5.45	9.19
σ_E^2	31.76	30.27	33.37

Table 3: Estimates for the two-way nested mixed effects model.

(c) The ICC and the total variability together with their confidence limits are given by:

Effect	Estimate	LCL	UCL
ICC	0.19	0.15	0.23
σ_T^2	39.08	36.82	41.57

(d)-(e) To investigate heteroschedasticity we run 2 different (nested models). The restricted model assumes equal between group variances whereas the "full" model assumes different between group variances. A LRT is performed to evaluate the null hypothesis of homoschedasticity (i.e. $H_0 = \sigma^2_{C(T)}(0) = \sigma^2_{C(T)}(1)$). Using the likelihood ratio test results in a test statistic of 1.0399 and 1 degree of freedom, and hence a p-value of 0.308. We therefore fail to reject the null-hypothesis and have thus not found enough evidence for heteroscedasticity.

Solution A8. We consider the following model:

$$\begin{aligned} &y_{hij} = \mu + \alpha_h + b_i + e_{hij}, \text{where:} \\ &\mu \text{is the over all mean of the observations,} \\ &\alpha_h \text{is the effect of sex, with} \alpha_2 = 0, \\ &b_i \sim \mathcal{N}(0, \sigma_C^2) \text{is the effect of class,} \\ &e_{hij} \sim \mathcal{N}(0, \sigma_R^2) \text{is the residual.} \end{aligned}$$

(a) Following the methods in the slides, we first determine the degrees of freedom of the included parameters: `CLASS` (C), `SEX` (S) and the residual $R = \text{Child}(SC)$. The degrees of freedom are: $I-1, H-1$ and $(J-1)HI$ respectively. The sums of squares then are:

$$\begin{aligned} SS_C &= \sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{\cdot i \cdot} - \bar{y}_{\cdot\cdot\cdot})^2 = \sum_{i=1}^{I} HJ(\bar{y}_{\cdot i \cdot} - \bar{y}_{\cdot\cdot\cdot})^2, \\ SS_S &= \sum_{h=1}^{H} IJ(\bar{y}_{h\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2, \\ SS_R &= \sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{hij} - \bar{y}_{hi\cdot})^2. \end{aligned}$$

The corresponding mean squares can be found by dividing the sums of squares by the degrees of freedom.

(b) The expected mean squares are given by:

$$\mathbb{E}[MS_R] = \sigma_R^2, \quad \mathbb{E}[MS_C] = \sigma_R^2 + HJ\sigma_C^2, \quad \mathbb{E}[MS_S] = \sigma_R^2 + IJQ_S,$$

where Q_S is a quadratic term for the fixed effects of sex.

(c) The resulting parameter estimates are:

$$\hat{\mu} = \bar{y}_{2..}, \quad \hat{\alpha}_h = \bar{y}_{h..} - \bar{y}_{2..}, \quad \hat{\sigma}^2_R = MS_R, \quad \hat{\sigma}^2_C = \frac{MS_C - MS_R}{HJ}.$$

For the data anlytical questions, we also consider the added interaction effect between `CLASS` and `SEX`.

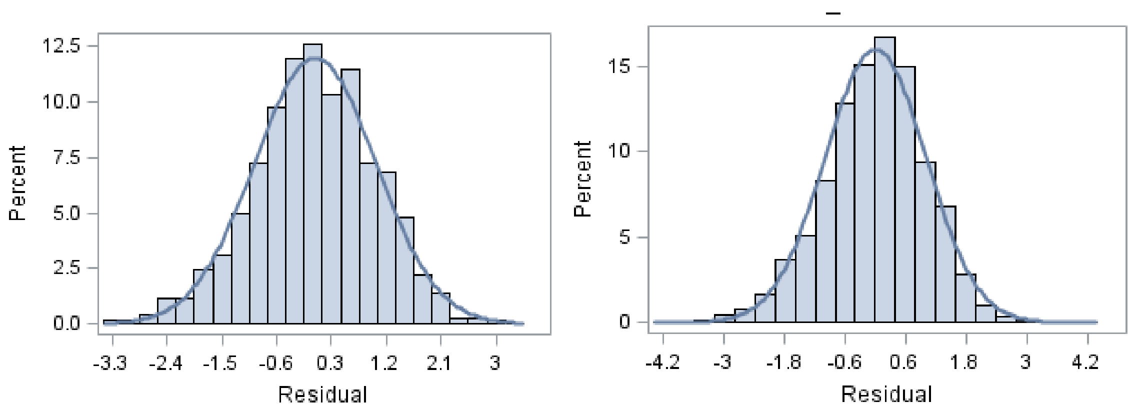
(d) The resulting estimates (with confidence limits) are given below:

Variable	Estimate	LCL	UCL
$\hat{\mu}$	7.37	6.88	7.85
$\hat{\alpha}_{SEX=0}$	-1.30	-1.71	-0.89
$\hat{\sigma}^2_C$	7.45	5.51	9.39
$\hat{\sigma}^2_{CS}$	0.58	-0.24	1.40
$\hat{\sigma}^2_R$	30.91	29.45	32.48

(e) Let ICC1 be the intraclass correlation coefficient of children of different sex in the same class and ICC2 be the intraclass correlation coefficient of children of the same sex and class. The estimates, with confidence limits are sumarized below:

Effect	Estimate	LCL	UCL
$ICC1$	0.19	0.15	0.23
$ICC2$	0.21	0.17	0.25
$\hat{\sigma}^2_T$	38.94	36.69	41.41

(f) The conditional and the marginal residuals (respectively) seem reasonably normal so the fit is okay (see picture below).



(a) Marginal residuals

(b) Conditional residuals

(g) A comparison of the four estimation methods is tabulized:

Quantity	TYPE1	TYPE3	ML	REML
$\hat{\mu}$	7.37 (6.88;7.85)	7.37 (6.88;7.85)	7.37 (6.88;7.86)	7.37 (6.88;7.86)
d.f.($\hat{\mu}$)	284	284	275	274
$\hat{\alpha}_{SEX=0}$	-1.30 (-1.71;-0.89)	-1.30 (-1.71;-0.89)	-1.30 (-1.70;-0.89)	-1.30 (-1.70;-0.89)
d.f.($\hat{\sigma}^2_C$)	162	162	169	168
$\hat{\sigma}^2_C$	7.45 (5.51;9.39)	7.45 (5.51;9.39)	7.60 (5.95;10.05)	7.64 (5.98;10.11)
$\hat{\sigma}^2_{CS}$	0.58 (-0.24;1.40)	0.58 (-0.24;1.40)	0.49 (0.16;6.88)	0.51 (0.17;6.25)
$\hat{\sigma}^2_R$	30.91 (29.44;32.48)	30.91 (29.44;32.48)	30.95 (29.49;32.52)	30.95 (29.49;32.52)

Solution A9.

(a) For the sum of squares we have:

$$\begin{aligned}
&\text{Sex}[H-1]: && SS_S = \sum_{h=1}^{H} IJK(\bar{y}_{h...} - \bar{y}_{....})^2 \\
&\text{Type}[I-1]: && SS_T = \sum_{i=1}^{I} HJK(\bar{y}_{.i..} - \bar{y}_{....})^2 \\
&\text{Class(Type)}[IJ-I]: && SS_{C(T)} = \sum_{i=1}^{I}\sum_{j=1}^{J} HK(\bar{y}_{.ij.} - \bar{y}_{.i..})^2 \\
&\text{Residual}[HIJK-HIJ]: && SS_R = \sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{hijk} - \bar{y}_{hij.})^2.
\end{aligned}$$

The mean squares can be found by dividing by the corresponding degrees of freedom (given in square brackets).

(b) We first write out the table:

	σ_S^2	σ_T^2	$\sigma_{C(T)}^2$	σ_R^2
S	X			X
T		X	X	X
$C(T)$			X	X
R				X

Then we fill in the sample sizes:

	σ_S^2	σ_T^2	$\sigma_{C(T)}^2$	σ_R^2
S	IJK			1
T		HJK	HK	1
$C(T)$			HK	1
R				1

And finally replace the VC's with quadratic functions for the fixed effects:

	Q_S	σ_T^2	$\sigma_{C(T)}^2$	σ_R^2
S	IJK			1
T		HJK	HK	1
$C(T)$			HK	1
R				1

Hence the expected mean squares are:

$$\begin{aligned}
\mathbb{E}[MS_S] &= Q_S + \sigma_R^2 \\
\mathbb{E}[MS_T] &= \text{HJK}\sigma_T^2 + \text{HK}\sigma_{C(T)^2} + \sigma_R^2 \\
\mathbb{E}[MS_{C(T)}] &= \text{HK}\sigma_{C(T)^2} + \sigma_R^2 \\
\mathbb{E}[MS_R] &= \sigma_R^2.
\end{aligned}$$

And consequently the estimators are:

$$\begin{aligned}
\hat{\sigma}_R^2 &= MS_R \\
\hat{\sigma}_{C(T)}^2 &= \frac{MS_{C(T)} - MS_R}{\text{HK}} \\
\hat{\sigma}_T^2 &= \frac{MS_T - MS_{C(T)}}{\text{HJK}} \\
\alpha_h &= \bar{y}_{h...} - \bar{y}_{....}.
\end{aligned}$$

(c) The methods for the previous problem still apply, which gives the same mean squares.

(d) The expected mean squares now become:

$$\begin{aligned}
\mathbb{E}[MS_S] &= \text{IJK}\sigma_S^2 + \sigma_R^2 \\
\mathbb{E}[MS_T] &= Q_T + \text{HK}\sigma_{C(T)^2} + \sigma_R^2 \\
\mathbb{E}[MS_{C(T)}] &= \text{HK}\sigma_{C(T)^2} + \sigma_R^2 \\
\mathbb{E}[MS_R] &= \sigma_R^2.
\end{aligned}$$

(e) And the estimators are:

$$\begin{aligned}
\hat{\sigma}_R^2 &= MS_R \\
\hat{\sigma}_{C(T)}^2 &= \frac{MS_{C(T)} - MS_R}{\text{HK}} \\
\hat{\sigma}_S^2 &= \frac{MS_S - MS_R}{\text{IJK}} \\
\alpha_i &= \bar{y}_{.i..} - \bar{y}_{....}.
\end{aligned}$$

Solution A10. We answer this question by fitting the full ANOVA model, that is we consider the model specified by the 3 main effects as well as their interactions. This will, namely, give us the opportunity to investigate different types of correlation coefficients of the ANOVA model.

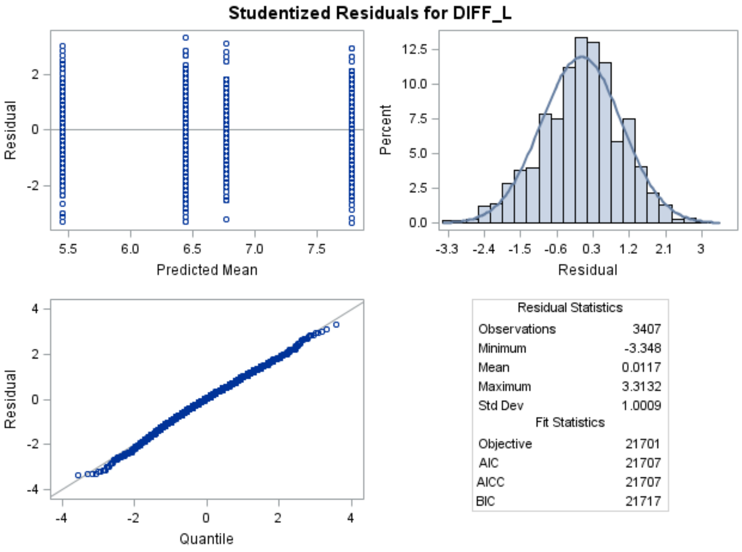
(a) The parameter estimates and their confidence intervals are:

Variable	Estimate	LCL	UCL
μ	6.76	5.88	7.65
$\alpha_{S=0}$	-1.32	-2.16	-0.47
$\alpha_{T=0}$	1.01	-0.0055	2.08
$\alpha_{S*T=(0,0)}$	-0.023	-1.00	0.96
$\sigma_{C(T)}^2$	7.02	5.10	8.93
$\sigma_{SC(T)}^2$	0.65	-0.22	1.52
σ_E^2	31.00	29.50	32.62

(b) Let ICC1 being the correlation of children of the same sex in one class ICC2 the correlation of the children of different sex in one class. The estimate of ICC1 and ICC2 together with the total variability and their confidence limits are given by:

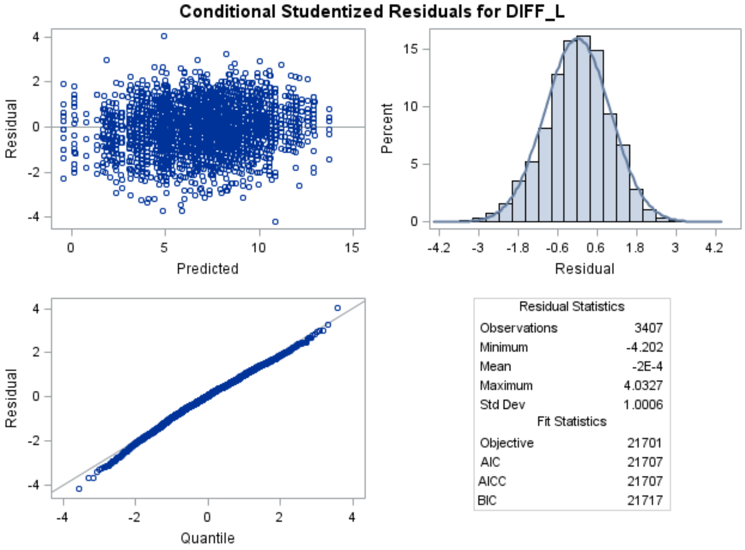
Effect	Estimate	LCL	UCL
$ICC1$	0.20	0.16	0.24
$ICC2$	0.18	0.14	0.22
$\hat{\sigma}_T^2$	38.66	36.41	41.13

(c) The conditional (studentized) residuals seem reasonably normal whereas the marginal (studentized) residuals slightly deviates from normality (see picture below):



Studentized Residuals for DIFF_L

Residual Statistics	
Observations	3407
Minimum	-3.348
Mean	0.0117
Maximum	3.3132
Std Dev	1.0009
Fit Statistics	
Objective	21701
AIC	21707
AICC	21707
BIC	21717



Conditional Studentized Residuals for DIFF_L

Residual Statistics	
Observations	3407
Minimum	-4.202
Mean	-2E-4
Maximum	4.0327
Std Dev	1.0006
Fit Statistics	
Objective	21701
AIC	21707
AICC	21707
BIC	21717

(d) The comparison of the 4 estimation methods is given in the table below:

Quantity	TYPE1	TYPE3	ML	REML
$\hat{\mu}$	6.77 (5.88;7.65)	6.77 (5.88;7.65)	6.76 (5.88;7.65)	6.76 (5.88;7.65)
d.f.($\hat{\mu}$)	381	386	373	367
$\hat{\alpha}_{S=0}$	-1.32 (-2.16;-0.47)	-1.32 (-2.16;-0.47)	-1.31 (-2.15;-0.48)	-1.31 (-2.15;-0.48)
d.f.($\hat{\alpha}_{S=0}$)	338	338	366	359
$\hat{\alpha}_{T=0}$	1.01 (-0.055;2.08)	1.01 (-0.053;2.08)	1.02 (-0.049;2.08)	1.02 (-0.054;2.09)
d.f.($\hat{\alpha}_{T=0}$)	327	331	319	315
$\hat{\alpha}_{S*T=(0,0}$	-0.023 (-1.00;0.96)	-0.022 (-1.00;0.96)	-0.028 (-0.99;0.94)	-0.026 (-1.00;0.94)
d.f.($\hat{\alpha}_{S*T=(0,0)}$)	249	249	268	264
$\hat{\sigma}^2_{C(T)}$	7.02 (5.10;8.93)	6.96 (5.07;8.86)	7.13 (5.53;9.55)	7.21 (5.58;9.67)
$\hat{\sigma}^2_{SC(T)}$	0.65 (-0.22;1.52)	0.65 (-0.22;1.52)	0.52 (0.17;6.61)	0.56 (0.19;5.74)
$\hat{\sigma}^2_{R}$	31.00 (29.50;32.62)	31.00 (29.50;32.62)	31.04 (29.54;32.65)	31.04 (29.54;32.65)

Solution A11. See the solution for A_9 for an extensive explanation.

(a)

$$
\begin{aligned}
&\text{Sex}[H-1]: && SS_S = \sum_{h=1}^{H} IJK(\bar{y}_{h...} - \bar{y}_{....})^2\\
&\text{Type}[I-1]: && SS_T = \sum_{i=1}^{I} HJK(\bar{y}_{.i..} - \bar{y}_{....})^2\\
&\text{Class(Type)}[IJ-I]: && SS_{C(T)} = \sum_{i=1}^{I}\sum_{j=1}^{J} HK(\bar{y}_{.ij.} - \bar{y}_{.i..})^2\\
&\text{Residual}[HIJK-HIJ]: && SS_R = \sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{hijk} - \bar{y}_{hij.})^2.
\end{aligned}
$$

(b) The expected mean squares are given by:

$$
\begin{aligned}
\mathbb{E}[MS_S] &= Q_S + \sigma^2_R\\
\mathbb{E}[MS_T] &= Q_T + \mathrm{HK}\sigma_{C(T)^2} + \sigma^2_R\\
\mathbb{E}[MS_{C(T)}] &= \mathrm{HK}\sigma_{C(T)^2} + \sigma^2_R\\
\mathbb{E}[MS_R] &= \sigma^2_R.
\end{aligned}
$$

(c) As a result, we have:

$$
\begin{aligned}
\hat{\sigma}^2_R &= MS_R\\
\hat{\sigma}^2_{C(T)} &= \frac{MS_{C(T)} - MS_R}{\mathrm{HK}}\\
\alpha_h &= \bar{y}_{h...} - \bar{y}_{....}\\
\beta_i &= \bar{y}_{.i..} - \bar{y}_{....}.
\end{aligned}
$$

Solution L1.

(a) We want to write

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j + \boldsymbol{e}_j,$$

where $\boldsymbol{y}_j$ is the a vector of repeated outcomes of unit class j, so that its entries correspond to all the outcomes for the children in class j. Thus, it has dimension $n_{11j} + n_{12j} + n_{21j} + n_{22j}$. The fixed-effects vector $\boldsymbol{\beta}$ is the 4-dimensional vector given by

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ (\alpha\beta)_{11} \end{pmatrix}.$$

The matrix $\boldsymbol{X}_j$ will then have a row corresponding to each children of class j. The first entry of each row will equal 1, the second entry will equal 1 if class j is of type 1 (otherwise 0), the third entry will equal 1 if the child has sex 1 (otherwise 0) and the last entry will equal 1 if the child has sex 1 and the class is of type 1 (otherwise 0).
Similarly, the random-effects vector $\boldsymbol{b}_i$ is the vector given by

$$\boldsymbol{b}_i = \begin{pmatrix} a_{j(i)} \\ a_{1j(i)} \\ a_{2j(i)} \end{pmatrix}.$$

The $\boldsymbol{Z}_j$ matrix will again have a row corresponding to each child of class j. The row will equal $(1, 1, 0)$ if the child has sex 1 hand $(1, 0, 1)$ if the child has sex 2.

(b) The $\boldsymbol{V}_j$ can be computed as

$$\boldsymbol{V}_j = \boldsymbol{Z}_j\boldsymbol{G}_j\boldsymbol{Z}_j^\top + \boldsymbol{R}_j,$$

where $\boldsymbol{R}_j$ is a matrix with diagonal entries σ_R^2 and 0 everywhere else, and

$$\boldsymbol{G}_j = \begin{pmatrix} \sigma^2_{C(T)} & 0 & 0 \\ 0 & \sigma^2_{SC(T)} & 0 \\ 0 & 0 & \sigma^2_{SC(T)} \end{pmatrix}.$$

Right-multiplication with a diagonal corresponds to simply multiplying each column of the original matrix with the corresponding diagonal entry. Thus, the entry $(\boldsymbol{V}_j)_{k_1,k_2}$ corresponding to children k_1 and k_2, is equal to the inner product of the k_1-th row of $\boldsymbol{Z}_j\boldsymbol{G}_j$ with the k_2-th row of $\boldsymbol{Z}_j$. Therefore, if $k_1 = k_2$, we will have

$$(\boldsymbol{V}_j)_{k_1,k_2} = \sigma^2_{C(T)} + \sigma^2_{SC(T)} + \sigma^2_R.$$

If $k_1 \neq k_2$ and both children have the same sex, then we have

$$(\boldsymbol{V}_j)_{k_1,k_2} = \sigma^2_{C(T)} + \sigma^2_{SC(T)}.$$

If $k_1 \neq k_2$ and the children have different sex, we will have

$$(\boldsymbol{V}_j)_{k_1,k_2} = \sigma^2_{C(T)}.$$

Solution L2. We use the following code:

```
PROC MIXED DATA=SCHOOLDATA METHOD=REML CL;
CLASS CLASS GIRL COMBI;
MODEL IQV= GIRL COMBI GIRL*COMBI/SOLUTION CL DDFM=SAT;
RANDOM GIRL/SUBJECT=CLASS GROUP=COMBI TYPE=UNR;
RUN;
```

The results are tabulated below. Remark that this problem results in numerical issues, as ρ_1 is estimated with a standard error of 0. The reason for this is that the estimation method maximizes the restricted likelihood, without taking constraints into consideration. In this case, the restricted likelihood was maximized for a correlation ¿1 and was therefore set to 1.

Effect	Estimate	Standard Error
μ	11.471	0.170
α_0	0.280	0.135
β_0	0.373	0.186
$(\alpha\beta)_{00}$	0.244	0.157
$\sigma_0^2(0)$	0.429	0.098
$\sigma_0^2(1)$	0.391	0.095
ρ_0	0.859	0.117
$\sigma_1^2(0)$	0.948	0.243
$\sigma_1^2(1)$	1.452	0.369
ρ_1	1	0
σ_R^2	3.779	0.090

Solution L3.

(a) The subject-specific model with linear time profiles can be written as (see slide 34):

$$y_{ij} = b_{i0} + b_{i1}t_{ij} + e_{ij},$$

where:

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2 \end{pmatrix}\right).$$

In specific, it states that for arbitrary $j, s \in \mathbb{N}$:

$$Cov(y_{ij}, y_{is}) = \tau_0^2 + \rho\tau_0\tau_1(t_{ij} + t_{is}) + t_{ij}t_{is}\tau_1^2,$$

in our case we set $t_{i1} = t, t_{i2} = (1+c)t$ and find:

$$Cov(y_{i1}, y_{i1}) = \tau_0^2 + \rho\tau_0\tau_1(2+c)t + (1+c)t^2\tau_1^2.$$

Furthermore we can see that:

$$\begin{aligned} Var(y_{i1}) &= \tau_0^2 + 2\rho\tau_0\tau_1 t + \tau_1^2 t^2 + \sigma_R^2 \\ Var(y_{i2}) &= \tau_0^2 + 2\rho\tau_0\tau_1(1+c)t + \tau_1^2(1+c)^2t^2 + \sigma_R^2, \end{aligned}$$

hence:

$$Cor(y_{i1}, y_{i2}) = \frac{\tau_0^2 + \rho\tau_0\tau_1(2+c)t + (1+c)t^2\tau_1^2}{\sqrt{\tau_0^2 + 2\rho\tau_0\tau_1 t + \tau_1^2t^2 + \sigma_R^2}\sqrt{\tau_0^2 + 2\rho\tau_0\tau_1(1+c)t + \tau_1^2(1+c)^2t^2 + \sigma_R^2}}.$$

(b) From the slides we know that:

$$\begin{aligned} \hat{\eta}(t)^2 &= Var(\hat{\beta}_0) + 2Cov(\hat{\beta}_0, \hat{\beta}_1)t + Var(\hat{\beta}_1)t^2; \\ \hat{\sigma}(t)^2 &= \hat{\tau_0}^2 + 2\hat{\rho}\hat{\tau}_0\hat{\tau}_1 t + t^2\hat{\tau}_1^2 + \hat{\sigma}_R^2, \end{aligned}$$

then we have that:

$$\begin{aligned} U(t) &= \hat{b}_{i0} + \hat{b}_{i1}t_{ij} + t_{df}^{-1}(1-\alpha)\sqrt{\hat{\sigma}^2 + \hat{\eta}^2}. \\ L(t) &= \hat{b}_{i0} + \hat{b}_{i1}t_{ij} - t_{df}^{-1}(1-\alpha)\sqrt{\hat{\sigma}^2 + \hat{\eta}^2}, \end{aligned}$$

with t_{df}^{-1} being the $1-\alpha$ quantile of the t-distribution with $df = m$ or $df = m - p$ (m is number of units and p is number of fixed effects) degrees of freedom and $\alpha = 0.025$.

(c) We use the following normalization for the follow up time (`FU`): (`TIME` − 120)/15. The results are tabulated below, using the notation from slide 35 and setting the placebo group as reference.

Parameter	Estimate	LCL	UCL
$\beta_{0,0}$	39.78	22.68	56.88
$\beta_{0,G}$	-4.74	-28.92	19.44
$\beta_{0,R}$	-1.83	-26.01	22.35
$\beta_{1,0}$	-1.15	-7.38	5.08
$\beta_{1,G}$	0.83	-7.98	9.64
$\beta_{1,R}$	-0.07	-8.88	8.74

Table 4: Fixed effects estimates for the model with subject-specific linear time profiles.

Parameter	Estimate	LCL	UCL
τ_0^2	462.30	253.16	1099.32
τ_1^2	49.37	23.76	157.56
ρ	-0.74	-0.99	-0.48
σ_R^2	112.09	77.95	174.94

Table 5: Covariance parameter estimates for the model with subject-specific linear time profiles.

(d) As the confidence limits for all treatment effects do contain 0, we fail to proof any treatment effects.

(e) Similar to the treatments, the confidence limits for the time effects contain 0.

(f) We've estimated the model by transforming the time axis ($FU = [t_{ij} - 120]/15$). Hence, we should estimate the correlation between observations at $FU = 1$ and $FU = 2$. From the slides,

$$\mathrm{COV}\left(y_{ij_1}, y_{ij_2}\right) = \tau_0^2 + \rho\tau_0\tau_1\left(t_{ij_1} + t_{ij_2}\right) + t_{ij_1}t_{ij_2}\tau_1^2.$$

Filling this in yields

$$\mathrm{COV}\left(y_{ij_1}, y_{ij_2}\right) = 462.30 - 0.738 * \sqrt{462.30}\sqrt{49.37}\,(1+2) + 2 * 49.37 \approx 226.6.$$

To obtain the correlation, we need to scale by the variances of y_{ij1} and y_{ij2}. Filling in the formulas under 3.1 yields $\widehat{Var}(y_{i1}) \approx 325$ and $\widehat{Var}(y_i2) \approx 400$. Scaling then yields for the correlation an estimate of 0.63.

(g) The mean time profiles can easily be found from the tables;

$$\begin{aligned} &\text{Placebo:} && \hat{\mu}(t) = 39.78 - 1.16t \\ &\text{G:} && \hat{\mu}(t) = 35.04 - .32t \\ &\text{R:} && \hat{\mu}(t) = 37.95 - 1.22t \end{aligned}$$

We now use the asymptotic covariance matrix for the fixed effect estimates and determine the estimated variance mean time profiles. We find

$$\hat{\eta}^2(t) = 67.60 - 36.27t + 8.97t^2.$$

Additionally we have $\hat{\sigma}^2(t) = 574.4 - 222.9t + 49.4t^2$. These can now be plugged in into the formulas for the reference limits and confidence limits for given t to obtain the 95% reference and confidence intervals.

E.g., at time 120 (implying $FU = 0$) the 95%-reference limits for placebo will be given by

$$39.78 \pm 2.08\sqrt{574.4 + 67.6}$$

and 95% confidence limits are given by

$$39.78 \pm 2.08 \cdot \sqrt{67.6}.$$

Solution L4. The code for this exercise is uploaded to canvas. We based our answers on the results from this code.

(a) Looking at the Type 3 tests for fixed effects we see that both `AGE*SEX` and `AGE_FP*SEX` is significant and hence the fractional polynomial is different per sex.

(b) Looking at the plot we see that the model indeed predicts the outcomes quite nicely.

(c) The normality is much better than we have seen before in the slides.

(d) See the code for extensive derivations.

Solution L5.

(a) The p-values of the Type3 tests are

Effect	pval
TRT	0.91
TIME	0.66
TRT*TIME	0.94

(b) The correlations appear to be quite strong among the time points. These are $\rho_{12} = 0.63$, $\rho_{13} = 0.64$, $\rho_{14} = 0.16$, $\rho_{23} = 0.74$, $\rho_{24} = 0.65$, $\rho_{34} = 0.45$.

(c) The differences are given in the table below.

Effect	Estimate	LCL	UCL
$(TRT_G - TRT_P)_{T=165}$	-2.32	-21.81	17.15
$(TRT_R - TRT_P)_{T=165}$	0.71	-18.77	20.19

(d) The differences are given in the table below.

Effect	Estimate	LCL	UCL
$(TRT_G)_{T=120} - 1/3\sum_{t=2}^{4}(\text{TRT}_G)_{T=t}$	-1.13	-16.52	14.25
$(TRT_R)_{T=120} - 1/3\sum_{t=2}^{4}(\text{TRT}_R)_{T=t}$	3.32	-12.06	18.71
$(TRT_P)_{T=120} - 1/3\sum_{t=2}^{4}(\text{TRT}_P)_{T=t}$	2.28	-13.11	17.66

Solution L6.

(a) The type 3 test for fixed effects indicate $p < 0.0001$ for `AGE*SEX` and $p = 0.0489$ for `AGE_FP*SEX` and thus it seems like there is a different polynomial per sex.

(b) We estimate the model using the `NOINT` option to simplify the calculations required. Assuming boys are 0 and girls are 1, we find:

$$\hat{\mu}_B(\mathtt{age}) = 2.7199 + 0.09939 \cdot \mathtt{age} - 0.4613 \cdot \mathtt{age}^{-0.5},$$
$$\hat{\mu}_G(\mathtt{age}) = 2.6581 + 0.1048 \cdot \mathtt{age} - 0.4753 \cdot \mathtt{age}^{-0.5}.$$

For the variance mean age profile, we get

$$\begin{aligned}\eta_B(\mathtt{age}) = {} & 0.000080 + \\ & 2 \cdot -5.8 \cdot 10^{-6} \cdot \mathtt{age} + 2 \cdot -0.00004 \cdot \mathtt{age}^{-0.5} + 2 \cdot 3.054 \cdot 10^{-6} \cdot \mathtt{age} \cdot \mathtt{age}^{-0.5} \\ & + 10^{-6} \cdot \mathtt{age}^2 \; + 0.000027 \cdot \mathtt{age}^{-1} \\ \eta_G(\mathtt{age}) = {} & 0.000069 + \\ & 2 \cdot -5.06 \cdot 10^{-6} \cdot \mathtt{age} + 2 \cdot -0.00003 \cdot \mathtt{age}^{-0.5} + 2 \cdot 2.671 \cdot 10^{-6} \cdot \mathtt{age}^{-0.5} \cdot \mathtt{age} + \\ & 8.779 \cdot 10^{-7} \cdot \mathtt{age}^2 + 0.000023 \cdot \mathtt{age}^{-1}\end{aligned}$$

The residual variance ($\hat{\sigma}(\texttt{age})$) is only estimated at the discrete times at which a subject's weight and age are recorded. Combining this information with the formulas on slide 68 allows for calculation of the reference limits. For reference limits at times where no measurement took place (e.g. 6 months), one needs a homoskedasticity assumption (i.e. $\sigma(\texttt{age}) = \sigma^2$).

Note that the sample size of the growth data set is quite large. For practical purposes, we could thus also use the asymptotic approximation.

(c) These reference limits are much narrower. This is because the marginal profile is estimating the sex-specific average profile, rather than an individual-specific average profile.

Solution L7. For the set-up of the model, look at A6. Now we need to look at different covariate structures with METHOD=REML. The values of the AICC and BIC can be found in the table below, from this you can see that UN is the structure that seems to fit the best.

Next you need to look at the importance of SEX in the model. The results can be seen in the following table, found using METHOD=ML. Here AGE and AGE05=AGE$^{-0.5}$ are always included in the model. Here it seems that all effects of SEX are needed in the model.

R	**UN**	AR(1)	TOEP	TOEP(H)	AR(H)
AICC	**-15981.3**	-14549	-14620	-	-14639
BIC	**-15680.8**	-14540	-14570	-	-14584

Table 6: AICC and BIC for different covariance structures

The model that seems to be the most reasonable is the model including the intercept and all interaction effects for SEX with AGE and AGE05, with an unstructered residual matrix.

	[INT+SEX+SEX*AGE+SEX*AGE05]	[INT+SEX]	[INT]
AICC	**-16029.5**	-15998	-15919
BIC	**-15701.7**	-15679	-15605

Table 7: AICC and BIC with/without sex