University of Glasgow

May 2020

EXAMINATION FOR THE DEGREES OF M.A., M.SCI. AND B.SC. (SCIENCE)

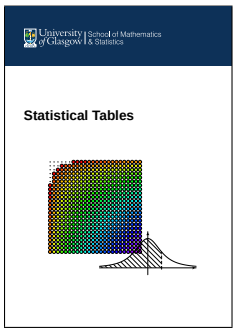
Statistics – Linear Mixed Models

This paper consists of 6 pages and contains 4 questions.
Candidates should attempt all questions.

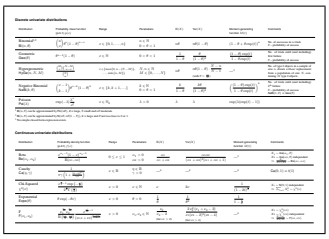
Question 1	20 marks
Question 2	20 marks
Question 3	20 marks
Question 4	20 marks
Total	80 marks

The following material is made available to you:

Statistical tables*



Probability formula sheet



"An electronic calculator may be used provided that it is allowed under the School of Mathematics and Statistics Calculator Policy. A copy of this policy has been distributed to the class prior to the exam and is also available via the invigilator."

CONTINUED OVERLEAF/

Candidates should attempt any three questions.

NOTE: If all four questions are attempted, candidates should clearly indicate which questions they wish to be marked. Otherwise, only the first three questions in the script book will be marked

1. A psychologist is interested in how the human brain responds to different colours of text. She decides to carry out an experiment to investigate whether the text colour affects people's ability to memorise a list of words. She proposes a simple memory test, where volunteers are given a set of randomly chosen words written in a particular colour and are given 5 minutes to try to memorise them. They then have 5 minutes to write down as many words as they can remember and are allocated a memory score based on their speed and accuracy. Each volunteer is asked to repeat the test three times on each colour (with a different set of words).

   (a) She recruits a set of 80 volunteers for this experiment, and wants to compare the memory scores obtained across four different colours (red, green, blue, black). She plans to ask each volunteer to carry out three memory tests for each of the four colours in turn. Write down a suitable model for this experiment, clearly defining the notation you use and stating all assumptions. **[6 MARKS]**

   (b) She realises that she does not have the time or resources to carry out that many memory tests and therefore has to carry out her experiment differently. Instead, she divides the volunteers into four groups of 20, and allocates each group to a different colour. Each volunteer is asked to carry out a single memory test using their chosen colour. Write down a suitable model for this experiment, clearly stating all assumptions. **[4 MARKS]**

   (c) Explain the difference between crossed and nested effects. Where possible use the experiments outlined in parts (a) and (b) to provide examples. **[2 MARKS]**

   (d) Explain the difference between balanced and unbalanced designs. Where possible use the experiments outlined in parts (a) and (b) to provide examples. **[2 MARKS]**

   (e) Specify the null hypothesis which the psychologist wishes to test. Describe this hypothesis with reference to the problem description, including notation from the model outlined in part (b). **[2 MARKS]**

   (f) We can use an F-test to test this null hypothesis. Write the formula for the test statistic for this F-test in terms of the relevant sums of squares (SSA etc) and any other appropriate quantities from the model outlined in part (b). **[4 MARKS]**

CONTINUED OVERLEAF/

2. A vet is interested in measuring the growth and development of baby rabbits, and recruits 206 bunnies for a study. The weight-for-age Z-score (WAZ) is a measure of relative weight, and can be used to monitor how well a rabbit is growing compared to their peers. Each rabbit has their WAZ score measured regularly during the first two months of their life. Let Y_{ij} be the WAZ score for rabbit i at timepoint j, and let x_{ij} be the age of rabbit i at timepoint j.

   (a) A model was fitted to this data using the following R code:

   ```
   mod1 <- lmer(WAZ ~ Age + ( Age | Rabbit ))
   ```

   Write down the model corresponding to this code, clearly defining the notation you use including distributional assumptions for any random terms in the model. **[6 MARKS]**

   (b) Determine $E(Y_{ij}|x_{ij})$ and $\text{Var}\big(Y_{ij}|x_{ij}\big)$ for this model. **[3 MARKS]**

   (c) The vet also considers three other models given by the following code:

   ```
   mod2 <- lmer(WAZ ~ Age + ( 1 | Rabbit ) + ( 0 + Age | Rabbit ))
   mod3 <- lmer(WAZ ~ Age + ( 1 | Rabbit ) )
   mod4 <- lmer(WAZ ~ Age)
   ```

   Explain how each of these three models differ from `mod1`. **[3 MARKS]**

   (d) We carry out a hypothesis test to compare models 1 and 2. The R output for this test is shown below.

   ```
   anova(mod1,mod2)

   mod2: WAZ ~ Age + ( 1 | Rabbit ) + ( 0 + Age | Rabbit )
   mod1: WAZ ~ Age + ( Age | Rabbit )

        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
   mod2  5 4277.9 4305.8 -2134.0   4267.9
   mod1  6 4212.6 4246.1 -2100.3   4200.6 67.305      1  2.325e-16 ***
   ```

      i. Using both notation and words, explain the hypothesis which is being tested here. **[2 MARKS]**

      ii. Which model would you choose, and why? **[2 MARKS]**

   (e) Partial output from fitting the preferred model is shown below.

CONTINUED OVERLEAF/

```
fixef(mod)

(Intercept)   Age
0.127282825 0.0559516

ranef(mod)


$Rabbit
      (Intercept)     Age
10001   0.6837219 -0.11054356
10002  -0.2583855 -0.06470521
10003   2.0106858 -0.14309790
10004   0.6325219 -0.10893392
```

i. Write out the estimated growth equation for the rabbit with id `10003`.
[2 MARKS]

ii. Predict the WAZ score for a new, unobserved rabbit at age 20 days.
[2 MARKS]

3. The general linear mixed model takes the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are design matrices for the fixed and random effects respectively and

$$E\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix},\ \text{Var}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}.$$

(a) This model can be rewritten as a multivariate normal distribution of the form

$$\mathbf{y} \sim N(\mathbf{M}, \mathbf{V}).$$

Write expressions for the mean vector $\mathbf{M}$ and the covariance matrix $\mathbf{V}$ in terms of the parameters of the general linear mixed model. **[3 MARKS]**

(b) Consider the simple linear mixed model

$$y_{ij} = \alpha_i + b_j + e_{ij}$$

CONTINUED OVERLEAF/

where α represents a fixed effect with two levels ($i = 1, 2$) and $\mathbf{b}$ represents a random effect with four levels ($j = 1, \ldots, 4$), where y_{ij} is defined for all combinations of i and j. Assume that $b_j \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$ independently of $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_E^2)$.

Express this model in the form of the general linear mixed model above using vector matrix notation. **[6 MARKS]**

(c) Write out the between-subject variance matrix $\mathbf{G}$, the error variance matrix $\mathbf{R}$ and the overall model variance matrix $\mathbf{V}$ in terms of σ_B^2 and σ_E^2. **[7 MARKS]**

(d) Explain the advantages of using restricted maximum likelihood (REML) instead of maximum likelihood (ML) when estimating the variance components in a mixed model. **[2 MARKS]**

(e) What does the abbreviation BLUP stand for in the context of parameter estimation in a mixed model? (Note that you do not have to explain what it means.) **[2 MARKS]**

4. A prominent UK-based professional cycling team decides to test new special 'dietary supplements' which they hope will enhance the performance of their athletes. Dave, their performance director, selects 20 cyclists from the team for his study. He asks 10 of them to take the new supplements while the other 10 continue to take the previous supplements. He compares their performance by measuring their maximal oxygen consumption (VO2 max) over a four month period. Each cyclist has their VO2 max measured at the start of the experiment and then each month for the next three months, giving four VO2 max measurements per cyclist (0, 1, 2, 3 months).

(a) A model was fitted to these data using the following R code:

```
mod1 <- gls(VO2 ~ Time*Supplement, data=cyclingPED,
            correlation = corSymm(form = ~ 1|Subject),
            weights=varIdent(form = ~ 1|Time))
```

Write down the mean model corresponding to the above code. Note from the model that time is treated as a factor rather than as continuous. **[5 MARKS]**

(b) The errors in this model are assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{R}$. Give the form that $\mathbf{R}$ takes when the following covariance structures are used.

i. unstructured (correlation = corSymm)

ii. compound symmetry (correlation = corCompSymm)

CONTINUED OVERLEAF/

[5 MARKS]

(c) Name another covariance structure which could be used. **[1 MARK]**

(d) The mean model in part (a) was fitted using each of the two covariance structures from part (b) above. We would like to carry out a likelihood ratio test to compare these two structures. Why is this appropriate for comparing these two models? **[2 MARKS]**

(e) The `-2 Res Log Lik` values for the models are as follows:

Unstructured: 730.8
Compound Symmetry: 763.7

Compute the test statistic for the likelihood ratio test. **[1 MARK]**

(f) What is the correct reference distribution for this likelihood ratio test? **[2 MARKS]**

(g) We obtain a p-value < 0.05 for our likelihood ratio test. What does this mean in terms of our covariance structure? **[2 MARKS]**

(h) What alternative approach could we use to compare covariance structures which are not nested? Provide detail on how the best structure would be selected under this approach. **[2 MARKS]**

END OF QUESTION PAPER.

University of Glasgow

May 2020
x

EXAMINATION FOR THE DEGREES OF M.A., M.SCI. AND B.SC. (SCIENCE)

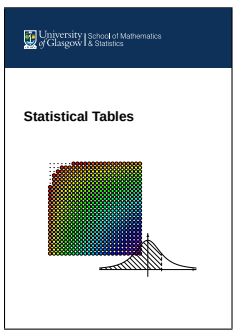
Statistics – Linear Mixed Models – Solutions

This paper consists of 8 pages and contains 4 questions.
Candidates should attempt all questions.

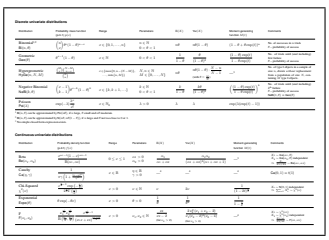
Question 1	0 marks
Question 2	0 marks
Question 3	0 marks
Question 4	0 marks
Total	0 marks

The following material is made available to you:

Statistical tables*



Probability formula sheet



"An electronic calculator may be used provided that it is allowed under the School of Mathematics and Statistics Calculator Policy. A copy of this policy has been distributed to the class prior to the exam and is also available via the invigilator."

CONTINUED OVERLEAF/

1. (a) Let Y_{ijk} be the score achieved on the ith colour by the jth volunteer on their kth attempt, with $i = 1, \ldots, 4$, $j = 1, \ldots, 80$ and $k = 1, \ldots, 3$.

Then

$$y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + e_{ijk}$$

- μ is the overall mean
- α_i is the fixed effect for the ith colour, $\sum_{i=1}^{4} \alpha_i = 0$
- b_j is the random effect for the jth volunteer, $b_j \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$
- $(\alpha b)_{ij}$ is the random effect for the interaction between colour and volunteer, $(\alpha b)_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_{AB}^2)$
- e_{ijk} are random errors, $e_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma_E^2)$

Here, b_j, $(\alpha b)_{ij}$, e_{ijk} are mutually independent random variables.

(b) Let Y_{ijk} be the score achieved on the ith colour by the jth volunteer, with $i = 1, \ldots, 4$, $j = 1, \ldots, 20$.

Then

$$y_{ijk} = \mu + \alpha_i + b_{j(i)}$$

- μ is the overall mean
- α_i is the fixed effect for the ith colour, $\sum_{i=1}^{4} \alpha_i = 0$
- $b_{j(i)}$ is the random effect for the jth volunteer, nested within colour i, $b_j \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$

[NOTE: There is no error term in the model because no replicates were taken.]

(c) The first model is an example of a **crossed** design, because each possible combination of factor A and factor B is present - every volunteer is tested on each colour.

The second model is an example of a **nested** design, because each volunteer is only tested on a single colour.

(d) Both of these experiments are examples of a **balanced** design, because each colour is tested by the same number of volunteers on the same number of occasions.

CONTINUED OVERLEAF/

An **unbalanced** design would be one where the number of volunteers or the number of times the experiment was carried out was different across the four colours.

(e) She wishes to test the hypothesis that all four colours are the same in terms of their average memory score.

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

(f) We can compute F as

$$F = \frac{MSA}{MSB(A)}$$

$$= \frac{SSA/(I-1)}{SSB/(N-I)} \qquad \text{where N = IJ}$$

$$= \frac{SSA/3}{SSE/76}$$

CONTINUED OVERLEAF/

2. (a) Let Y_{ij} be the WAZ score for rabbit i at age j.

Then

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij} + e_{ij}$$

- β_0 is a fixed effect representing the population intercept.
- β_1 is a fixed effect representing the population slope.
- $b_{0i} \overset{\text{iid}}{\sim} N(0, \sigma_0^2)$ is the intercept random effect for rabbit i
- $b_{1i} \overset{\text{iid}}{\sim} N(0, \sigma_1^2)$ is the slope random effect for rabbit i
- $\text{Corr}\big(b_{0i}, b_{1i}\big) = \rho \neq 0$.
- $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_E^2)$ is the error term.

Random variables b_{0i}^* are independent of e_{ij} and random variables b_{1i}^* are independent of e_{ij}.

(b)

$$E(Y_{ij}|x_{ij}) = \beta_0 + \beta_1 x_{ij}$$

$$\begin{aligned} \text{Var}\big(Y_{ij}|x_{ij}\big) &= \text{Var}\big(b_{0i}\big) + \text{Var}\big(b_{1i}x_{ij}\big) + 2\,\text{Cov}\big(b_{0i}, b_{1i}x_{ij}\big) + \text{Var}\big(e_{ij}\big) \\ &= \sigma_0^2 + \sigma_1^2 x_{ij}^2 + 2\rho\sigma_0\sigma_1 x_{ij} + \sigma_E^2 \end{aligned}$$

(c) Model 2 assumes uncorrelated intercept and slope random effects ($\rho = 0$).
Model 3 assumes that all rabbits have a common slope (ie $b_{1i} = 0$ for all i).
Model 4 assumes that all rabbits have a common slope and intercept (ie a normal linear model).

(d) i.

$$H_0 : \text{Corr}\big(b_{0i}, b_{1i}\big) = 0$$

We are testing the hypothesis that the slope and intercept random effects are not correlated.

ii. We obtain a p-value < 0.05, and therefore reject the null hypothesis. We should use the model with the correlated random effects (model 1).

(e) i.

$$\begin{aligned} Y_{ij} &= 0.127 + 0.056x_{ij} + 2.011 + (-0.143)x_{ij} \\ &= 2.138 - 0.087x_{ij} \end{aligned}$$

CONTINUED OVERLEAF/

ii. We can only used the fixed part of the model

$$Y = 0.127 + 0.056 * 20 = 1.247$$

3. (a) $\mathbf{M} = \mathbf{X}\boldsymbol{\beta}$

$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\intercal} + \mathbf{R}$

(b) First define the model components

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \end{bmatrix}; \ \boldsymbol{\beta} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}; \ \mathbf{u} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}.$$

We then create the corresponding design matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}; \ \mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

(c) $\mathbf{G} = \sigma_B^2 \mathbf{I}_4$
$\mathbf{R} = \sigma_E^2 \mathbf{I}_8$
This gives us a combined variance matrix for the model of:

$$\mathbf{V} = \begin{bmatrix} \sigma_B^2 + \sigma_E^2 & 0 & 0 & 0 & \sigma_B^2 & 0 & 0 & 0 \\ 0 & \sigma_B^2 + \sigma_E^2 & 0 & 0 & 0 & \sigma_B^2 & 0 & 0 \\ 0 & 0 & \sigma_B^2 + \sigma_E^2 & 0 & 0 & 0 & \sigma_B^2 & 0 \\ 0 & 0 & 0 & \sigma_B^2 + \sigma_E^2 & 0 & 0 & 0 & \sigma_B^2 \\ \sigma_B^2 & 0 & 0 & 0 & \sigma_B^2 + \sigma_E^2 & 0 & 0 & 0 \\ 0 & \sigma_B^2 & 0 & 0 & 0 & \sigma_B^2 + \sigma_E^2 & 0 & 0 \\ 0 & 0 & \sigma_B^2 & 0 & 0 & 0 & \sigma_B^2 + \sigma_E^2 & 0 \\ 0 & 0 & 0 & \sigma_B^2 & 0 & 0 & 0 & \sigma_B^2 + \sigma_E^2 \end{bmatrix}$$

CONTINUED OVERLEAF/

(d) Restricted maximum likelihood accounts for the degrees of freedom lost in estimating the fixed effect parameters. Maximum likelihood does not.

(e) Best Linear Unbiased Predictor.

CONTINUED OVERLEAF/

4. (a) The mean model for these data is

$$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

- μ is the overall mean
- α_i is the fixed effect for supplement with $i = 1, 2$.
- β_j is the fixed effect for time in months with $j = 1, 2, 3, 4$.
- $(\alpha\beta)_{ij}$ is the interaction between supplement and month.

(b) Matrix $\mathbf{R}$ is block-diagonal with 20 blocks, each corresponding to an individual cyclist. Under this model the blocks are identical for all cyclists.

i. Unstructured covariance structure: each block is of the form

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_3^2 & \sigma_{34} \\ & & & \sigma_4^2 \end{bmatrix}$$

where σ_{12} is the covariance between finishing times in week 1 and 2 etc.

ii. Compound symmetry: each block is of the form

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_B^2 + \sigma_E^2 & \sigma_B^2 & \sigma_B^2 & \sigma_B^2 \\ & \sigma_B^2 + \sigma_E^2 & \sigma_B^2 & \sigma_B^2 \\ & & \sigma_B^2 + \sigma_E^2 & \sigma_B^2 \\ & & & \sigma_B^2 + \sigma_E^2 \end{bmatrix}$$

Note: Either of the two versions of the matrix for compound symmetry gets full marks.

(c) Any sensible answer (eg Toeplitz, AR(1), exponential) will be accepted.

(d) These two structures are nested within each other - we can obtain the compound symmetry structure by fixing all the parameters in the unstructured matrix to be equal to each other.

(e) TS $= 763.7 - 730.8 = 32.9$

CONTINUED OVERLEAF/

(f) $\chi^2(8)$, since the unstructured matrix has 8 more parameters than the compound symmetry one.

(g) There is a significant difference between the models fitted under the two structures. An unstructured correlation matrix is necessary.

(h) We could use information criteria (eg AIC, BIC) and select the model which produces the minimum value for the chosen criterion.

END OF QUESTION PAPER.