University of Glasgow

May 2018

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

Statistics – Linear Mixed Models

"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

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 statistics lecturer has two pet rabbits who absolutely love eating hay. There are three main types of hay available for rabbits - alfalfa hay, meadow hay and timothy hay. He decides to carry out an experiment to investigate which of these types of hay rabbits prefer. A total of 10 rabbits are recruited to the study, and each rabbit is given each type of hay for a total of three days. The order in which the rabbits are fed the types of hay is randomly decided in advance. Each day, we measure the total weight of hay (in grams) eaten by each rabbit, giving a total of 90 observations (10 rabbits, 9 days).

   (a) Write down a suitable model for this experiment, clearly stating all assumptions. **[6 MARKS]**

   (b) With reference to this example, explain the difference between a nested and a crossed experiment. **[2 MARKS]**

CONTINUED OVERLEAF/

(c) Using the sums of squares below, test the hypothesis that there is no difference between the types of hay in terms of preference. State your conclusion both in terms of the model parameters and in relation to the original example.

Source	SS
hay	93192
rabbit	26136
hay:rabbit	67301
error	8641

One or more of the following distributional results may be useful:

$F(2, 9; 0.95) = 4.26$ $F(2, 18; 0.95) = 3.55$ $F(2, 60; 0.95) = 3.15$
$F(3, 9; 0.95) = 3.86$ $F(3, 10; 0.95) = 3.71$ $F(3, 90; 0.95) = 2.70$

[6 MARKS]

(d) Provide unbiased (method of moments) point estimates for each of the variance components in your model. [Note that it is possible for these estimates to be negative] **[6 MARKS]**

2. Suppose we have data (x_{ij}, Y_{ij}) where i indicates a grouping/clustering level and j indicates measurements within groupings. We consider a model of the form:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b^*_{0i} + b^*_{1i} x_{ij} + e_{ij}$$

where β_0, β_1 are unknown parameters; b^*_{0i}, b^*_{1i} are random effects; and e_{ij} is the error term.

(a) This is known as a *random coefficient* model. Explain how it differs from an *analysis of covariance (ANCOVA).* **[2 MARKS]**

(b) Write down all of the distributional assumptions for e_{ij}, b^*_{0i} and b^*_{1i} in the context of this random coefficient model. **[4 MARKS]**

(c) Determine $E(Y_{ij}|x_{ij})$ and $\text{Var}\big(Y_{ij}|x_{ij}\big)$. **[2 MARKS]**

(d) Creatine is a supplement widely used by bodybuilders and athletes to build muscle strength. In a study, 42 subjects were were asked to take 7 grams of creatine per day over a 5-week period. The subjects had their muscle strength measured at the start of the study and then again every week until the end of the study, giving a total of 6 measurements per subject.

The following four models were fitted to this study data using the `lme4` library in R.

CONTINUED OVERLEAF/

```
m1 <- lmer(Muscle ~ Time + (Time|Subject))
m2 <- lmer(Muscle ~ Time + (1|Subject)+(0+Time|Subject))
m3 <- lmer(Muscle ~ Time + (1|Subject))
m4 <- lm(Muscle ~ Time)
```

Define each of these four models, both in terms of the creatine example **and** the normal linear model described in part (a). You may wish to use sketches and/or make reference to the model parameters. **[6 MARKS]**

(e) The following model comparisons were carried out:

```
anova(m1,m2,m3,m4)

#This compares models sequentially (m4 v m3), (m3 v m2), (m2 v m1)

Models:
m4: Muscle ~ Time
m3: Muscle ~ Time + (1 | Subject)
m2: Muscle ~ Time + (1 | Subject) + (0 + Time | Subject)
m1: Muscle ~ Time + (Time | Subject)
   Df    AIC    BIC   logLik deviance   Chisq Chi Df Pr(>Chisq)
m4  3 2041.8 2052.4 -1017.91   2035.8
m3  4 1964.1 1978.2  -978.03   1956.1 79.7474      1  < 2.2e-16 ***
m2  5 1956.1 1973.8  -973.06   1946.1  9.9490      1   0.001609 **
m1  6 1958.0 1979.2  -973.01   1946.0  0.0891      1   0.765376
```

Based on these results, which model would you select, and why? **[2 MARKS]**

(f) Selected output from the final correct model is shown below.

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  49.4249     1.6648  29.688
Time          4.1103     0.4612   8.911

> ranef()
$Subject
    (Intercept)       Time
1    1.39676209  1.8205717
2    4.20886279  1.9158831
3   -3.94390531 -2.2711469
4   -6.28764761 -1.1444901
5   12.48438416 -0.2981106
6    7.56926501  1.9019373
7    0.84347632  1.1823382
```

CONTINUED OVERLEAF/

```
8   -1.73151497  0.1020777
...
```

i. Predict the muscle strength of Subject 6 after 4 weeks.

ii. Predict the muscle strength of a new, unobserved subject in week 3.

[4 MARKS]

3. A cancer research charity is interested in investigating some of the factors which determine whether or not cancer treatment will be successful for patients. A researcher obtained data for 2000 lung cancer patients who had been given treatment in US hospitals - these patients are treated with chemotherapy and have their cancer progress monitored monthly over a 6-month period. The response variable for the study is **success** which is coded as 0 if the patient still has cancer symptoms and 1 if the patient is in remission (ie has no symptoms of cancer). The dataset also contains a number of explanatory variables: **sex** (0=female, 1=male); **age** - a continuous variable measuring age in years, **tumour** - the size of the patient's cancerous tumour (in cm), **month** - the month of the treatment (1-6), **id** - a unique patient id.

(a) The researcher decides to fit a generalised linear mixed model (GLMM). What feature of this dataset makes a GLMM a more appropriate choice than a standard mixed model? **[2 MARKS]**

(b) Another alternative model for this type of data would be a generalised estimating equation (GEE). What are the main differences between a GLMM and a GEE? **[2 MARKS]**

(c) In a GLMM, the conditional mean relates to the linear predictor through a link function:

$$g(E(\mathbf{y}|\mathbf{u})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}.$$

What would be an appropriate link function $g()$ for this example? **[2 MARKS]**

(d) A model was fitted to these data in R. Selected output is shown below.

```
mod1 <- glmer(success ~ sex + age + tumour + month + ( 1 | id ))

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.52166    1.06991   1.422  0.15496
sexMale      0.66822    0.21086   3.169  0.00153 **
age          0.01449    0.01419   1.021  0.30731
tumour      -1.44060    0.26811  -5.373 7.74e-08 ***
```

CONTINUED OVERLEAF/

```
month        0.39948    0.03242  12.322  < 2e-16 ***


Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 1.31     1.145
Number of obs: 1046, groups:  id, 356

ranef(mod1)
$id
     (Intercept)
1    1.026738316
2    0.045266159
3    0.552361630
4    0.265157840
5   -0.511174643
...
```

Explain the fixed effects output in terms of each of the covariates in our model. **[6 MARKS]**

(e) In a standard mixed model, we would test the significance of the random effect term by comparing the models with and without the random effect. Why can't we do this here? **[2 MARKS]**

(f) Subject 1 is a 45 year-old male with a 3cm tumour. What is the probability that his treatment will have been successful after 5 months? **[6 MARKS]**

4. A psychologist is interested in how quickly young children can learn to complete cognitive tasks. 50 children were recruited for the study, and she measured how long it took each child to complete a card matching task. The children were asked to repeat the task on five consecutive days, to see how much their performance had improved over time. She was also interested in whether boys and girls performed differently.

   The dataset contains the following variables: **id** - a unique indentifier for each child; **time** - the time in seconds for the child to complete the task; **day** - the day of the experiment (1-5); **sex** - the sex of the child (0=female, 1=male).

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

```
mod1 <- gls(time ~ day*sex, data=memory,
            correlation = corSymm(form = ~ 1|id),
            weights=varIdent(form = ~ 1|day))
```

CONTINUED OVERLEAF/

Write down the mean model corresponding to this code. **[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. AR1 (correlation = corAR1)

[5 MARKS]

(c) The mean model in part (a) was fitted using each of the two covariance structures above. We would like to carry out a likelihood ratio test to compare these two structures.

i. Explain why a likelihood ratio test is appropriate for comparing these structures.

[2 MARKS]

ii. Explain why the asymptotic reference distribution is appropriate for this likelihood ratio test.

[2 MARKS]

iii. What is the correct reference distribution for this likelihood ratio test?

[2 MARKS]

iv. A likelihood ratio test was carried out, and we obtained the results below. Note that `mod1` is unstructured and `mod2` is AR1.

```
anova(mod1, mod2)
     Model df      AIC      BIC    logLik   Test L.Ratio p-value
mod1     1 19 1897.502 1964.104 -929.7512
mod2     2  6 1901.268 1922.300 -944.6341 1 vs 2 29.7659  0.0051
```

Explain what the result of this likelihood ratio test means in terms of our covariance structure. **[2 MARKS]**

v. The AIC and the BIC appear to disagree with each other in terms of the preferred model. Explain why this might occur. **[2 MARKS]**

END OF QUESTION PAPER.

University of Glasgow

May 2018
x

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

Statistics – Linear Mixed Models – Solutions

"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

1. **[Q1 is an unseen data example.]**

   (a) Let Y_{ijk} be the amount of the ith type of hay eaten by the jth rabbit on the kth day, with $i = 1, \ldots, 3$, $j = 1, \ldots, 10$ and $k = 1, \ldots, 3$.

   [1 MARK]

   Then

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

   [1 MARK]

   - μ is the overall mean
   - α_i is the fixed effect for the ith type of hay, $\sum_{i=1}^{3} \alpha_i = 0$
   - b_j is the random effect for the jth rabbit, $b_j \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$

CONTINUED OVERLEAF/

- $(\alpha b)_{ij}$ is the random effect for the interaction between hay type and rabbit, $(\alpha b)_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_{AB})$

- e_{ijk} are random errors, $e_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2_E)$

Here, b_j, $(\alpha b)_{ij}$, e_{ijk} are mutually independent random variables. **[4 MARKS]**

(b) This is an example of a **crossed** design, because each possible combination of factor A and factor B is present - that is, every rabbit eats each type of hay.

A **nested** design is when each element of factor A only appears in conjunction with one particular element of factor B. This would apply if each the rabbits were split into three groups and each group only received a single type of hay.

[2 MARKS]

(c) We wish to test the hypothesis: $H_A : \alpha_1 = \alpha_2 = \alpha_3$. **[1 MARK]**

The test statistic is given by

$$F_A = \frac{MSA}{MSAB} = \frac{SSA/(I-1)}{SSAB/(I-1)(J-1)} \frac{93192/2}{67301/18} = 12.46.$$

[2 MARKS]

This is greater than the critical value $F(2, 18; 0.95) = 3.55$, and therefore we reject the null hypothesis and conclude that there is a significant difference in preference between the types of hay. **[3 MARKS]**

(d) We have MSB = 26136/9 = 2904, MSAB = 67301/18 = 3738.94 and MSE = 8641/60 = 144.02. We use these to obtain estimates for σ^2_E, σ^2_{AB} and σ^2_B as follows:

$$E(MSE) = \sigma^2_E \Rightarrow \hat{\sigma}^2_E = MSE = 144.02$$

$$E(MSAB) = \sigma^2_E + K\sigma^2_{AB} = \sigma^2_E + 3\sigma^2_{AB}$$
$$\Rightarrow \hat{\sigma}^2_{AB} = \frac{MSAB - \hat{\sigma}^2_E}{3} = \frac{3738.94 - 144.02}{3} = 1198.31$$

$$E(MSB) = \sigma^2_E + K\sigma^2_{AB} + IK\sigma^2_B = \sigma^2_E + 3\sigma^2_{AB} + 9\sigma^2_B$$
$$\Rightarrow \hat{\sigma}^2_B = \frac{MSB - 3\hat{\sigma}^2_{AB} - \hat{\sigma}^2_E}{9} = \frac{MSB - MSAB}{9} = -92.77$$

[6 MARKS]

CONTINUED OVERLEAF/

2. **[Q2(a)-(c) are bookwork, (d)-(f) are an unseen data example.]**

(a) In a random coefficient model, the regression coefficients for continuous explanatory variables are assumed to be random effects. In ANCOVA models they were assumed to be fixed effects. **[2 MARKS]**

(b)
- $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_E^2)$
- $b_{0i}^* \overset{\text{iid}}{\sim} N(0, \sigma_0^2)$
- $b_{1i}^* \overset{\text{iid}}{\sim} N(0, \sigma_1^2)$
- $\text{Corr}\big(b_{0i}^*, b_{1i}^*\big) = \rho \neq 0.$

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

[4 MARKS]

(c)

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

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

[2 MARKS]

(d)
- m1 is a model with a random slope and a random intercept, with correlation between the slope and intercept. The starting strength and change in strength over time can be different for each individual bodybuilder.
- m2 is the same as m1 except that the slope and intercept effects are assumed to be independent.
- m3 is a model with a random intercept only. Each bodybuilder may have a different starting strength, but the change in strength over time is assumed to be the same for all individuals.
- m4 is a linear regression with no random effects. The starting strength and change in strength over time are assumed to be the same for all bodybuilders.

[6 MARKS]

(e) We would pick m2, because the likelihood ratio test suggests no difference between m2 and m1, and therefore we select the model with fewer parameters (m2). Note that m2 also has the lowest AIC and BIC. **[2 MARKS]**

(f) i. The best prediction for the ith subject is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{ij} + BLUP(b_{0i}) + BLUP(b_{1i})x_{ij}$$

CONTINUED OVERLEAF/

For $i = 6$ and $x_{ij} = 4$ this equals

$$49.4249 + 4.1103 \times 4 + 7.5693 + 1.9019 \times 4 = 81.043$$

ii. The best prediction for an unobserved subject is given by the fixed effect terms in the model. This is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_{ij} = 49.4249 + 4.1103 \times 3 = 61.7558.$$

[4 MARKS]

3. **[Q3(a)-(c), (e) are bookwork, (d) and (f) are an unseen data example.]**

(a) Standard mixed models require a continuous response. Here, our response is binary and therefore a GLMM is more suitable. **[2 MARKS]**

(b) GLMM allows random effects to be incorporated, GEE does not. GEE focuses on the impact of covariates on the mean response for the population. GLMM allows you to address the impact of covariates on the mean response for an individual. **[2 MARKS]**

(c) We have a binary response, therefore $g()$ should be a logit function. [NB: 1 mark for reference to binomial] **[2 MARKS]**

(d)
- There is a significant sex effect in this model. The odds of success for males are $\exp(0.66822) = 1.95$ times higher than for females.
- There is no significant age effect in the model. Age does not appear to affect the chances of success.
- There is a significant tumour effect in the model. The odds of success are multiplied by $\exp(-1.44060) = 0.237$ for every 1cm increase in the size of the tumour.
- There is a significant month effect in the model. The odds of success are multiplied by $\exp(0.39948) = 1.491$ for every month of treatment.

[6 MARKS]

(e) By definition, a GLMM must contain random effects, and therefore we cannot fit the model without them in order to make a comparison. **[2 MARKS]**

(f) The probability of success for Subject 1, a 45 year-old male with a 3cm tumour in month 5 is given by:

$$\frac{\exp(1.52166 + 0.66822 + 0.01449 \times 45 + (-1.44060) \times 3 + 0.39948 \times 5 + 1.0267)}{1 + \exp(1.52166 + 0.66822 + 0.01449 \times 45 + (-1.44060) \times 3 + 0.39948 \times 5 + 1.0267)} = 0.824$$

CONTINUED OVERLEAF/

There is an 82.4% chance that this patient's treatment will have been successful after 5 months. **[6 MARKS]**

4. **[Q4 is an unseen data example. Parts (b), (c)(i, ii, v) are bookwork]**

(a) The mean model for these data is

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

[2 MARKS]

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

[3 MARKS]

(b) Matrix $\mathbf{R}$ is block-diagonal with 50 blocks, each corresponding to an individual child. Under this model the blocks are identical for all children.

[1 MARK]

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

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ & & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ & & & \sigma_4^2 & \sigma_{45} \\ & & & & \sigma_5^2 \end{bmatrix}$$

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

[2 MARKS]

ii. AR(1): each block is of the form

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & 1 & \rho & \rho^2 & \rho^3 \\ & & 1 & \rho & \rho^2 \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$$

CONTINUED OVERLEAF/

where ρ is the correlation between measurements taken 1 day apart and measurements taken k weeks apart have correlation ρ^k

[2 MARKS]

(c) i. A likelihood ratio test is appropriate because two structures being compared are **nested** - the AR(1) structure can be obtained by fixing the values of parameters in the unstructured model. **[2 MARKS]**

ii. The asymptotic reference distribution can be used as long as we are not fixing a parameter equal to a boundary constraint. Correlation parameters can take any value in the range [-1,1], and we are not fixing any parameters to these values. **[2 MARKS]**

iii. $\chi^2(13)$, since the unstructured matrix has 13 more parameters than the AR. **[2 MARKS]**

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

[2 MARKS]

v. These two criteria are calculated differently. BIC tends to prefer less complex models than AIC. **[2 MARKS]**

END OF QUESTION PAPER.