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 commercial baker wants to develop the perfect lemon sponge cake. The perfect cake should be fluffy and moist, and should produce a nice clean slice when you cut into it. A cake which is too dry will crumble when cut, while a cake which is too moist will stick to the knife, both of which result in the loss of cake. The quality of the cake can therefore be measured by the mass of cake loss (in grams) when cut, with a loss of 0 grams representing a perfect cake. She wishes to choose between two recipes (A and B) to identify which produces the lowest amount of cake loss. She randomly selects 4 of her chefs to take part in the experiment, and each chef is asked to cook 5 cakes using each recipe. Each of the 40 cakes is then cut and the cake loss is measured.

   The dataset contains the variables `loss` (g of cake lost), `recipe` (Recipe A or B) and `chef` (chef who cooked the cake, taking values 1, 2, 3, 4).

CONTINUED OVERLEAF/

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

(b) This model was fitted, and we obtained the table below.

Source	DF	SS
recipe	I-1	5831
chef	J-1	434.58
recipe*chef	(I-1)(J-1)	401.29
error	IJ(K-1)	1032

Test the hypothesis that there is no difference between her two recipes in terms of cake loss. State your conclusion in terms of the model parameters, and then relate this back to the practical example.

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

$F(3, 16; 0.95) = 3.24$ $\quad F(3, 1; 0.95) = 215.71$ $\quad F(2, 4; 0.95) = 6.94$

$F(1, 3; 0.95) = 10.13$ $\quad F(24, 1; 0.95) = 249.05$ $\quad F(4, 2; 0.95) = 19.25$

[5 MARKS]

(c) Provide method of moments point estimates for each of the three variance components in the model. The following equations may be useful:

$$E(MSB) = \text{Var(error)} + 5\ \text{Var(recipe*chef)} + 10\ \text{Var(chef)}$$
$$E(MSAB) = \text{Var(error)} + 5\ \text{Var(recipe*chef)}$$

[4 MARKS]

(d) Let δ represent the mean difference in cake loss between Recipe A and Recipe B. Derive an expression for the variance of δ. **[3 MARKS]**

(e) With reference to this example, explain the difference between **crossed** and **nested** factors.

[2 MARKS]

2. Scientists carried out an experiment to test whether a new protein supplement could improve athletic performance. A group of 50 amateur runners were recruited; 25 were given a regular four week course of the new supplement, while the remaining athletes were given a four week course of a protein-free supplement (placebo). Prior to starting the experiment, each athlete was asked to complete a 5km run, and their baseline time was recorded. Each week during the experiment, they did another timed 5km run to monitor their progress.

CONTINUED OVERLEAF/

(a) A model was fitted to this data using the following SAS code.

```
proc mixed;
class id week supp;
model runtime = basetime week supp week*supp;
repeated week / type=un subject=id;
run;
```

Write down the mean model corresponding to the above 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 each of the following covariance structures are used.

i. unstructured (type=un)

ii. AR(1) (type=ar)

iii. compound symmetry (type=cs)

[8 MARKS]

(c) The mean model in part (a) was fitted with the three different covariance structures mentioned. Based on the fit statistics below, which covariance structure(s) are the best candidates? If you have more than one candidate structure, carry out a formal test to compare them.

	un	ar(1)	cs
-2 Res Log Likelihood	762.1	784.2	791.7
AIC (Smaller is Better)	782.1	786.2	793.7
AICC (Smaller is Better)	783.4	786.3.	793.8
BIC (Smaller is Better)	793.1	790.2.	791.8

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

$\chi^2(8; 0.95) = 15.5 \qquad \chi^2(9; 0.95) = 16.9 \qquad \chi^2(10; 0.95) = 18.3.$

[4 MARKS]

(d) 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 given matrices and

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

CONTINUED OVERLEAF/

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

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

i. Complete the above expression by providing the correct values for θ and λ.

ii. Write an expression for the overall variance $\mathbf{V}$ in terms of the between-subject variance $\mathbf{G}$ and the error variance $\mathbf{R}$.

[3 MARKS]

3. Suppose we have data (x_{ij}, Y_{ij}) (where i labels subjects) for which we consider the model for Y_{ij} given x_{ij} of the form

$$M_0 : Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij} + e_{ij},$$

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

(a) i. Write down the distributional assumptions for e_{ij}, b_{0i} and b_{1i} in the context of a normal linear mixed model with (possibly correlated) random coefficients.

ii. Determine $E(Y_{ij}|x_{ij})$ and $\text{Var}\big(Y_{ij}|x_{ij}\big)$ under these assumptions.

[4 MARKS]

(b) Identify the following three special cases in terms of the covariance parameters in the model.

i. M_1: ‘independent random intercept and slope for each subject’;

ii. M_2: ‘random intercept for each subject’; and

iii. M_3: ‘no random subject effects’.

[3 MARKS]

(c) Provide a rough sketch of some subject-specific regression lines to illustrate the differences between models M_1, M_2 and M_3.

[3 MARKS]

(d) The diameters of 12 apples on a particular tree were measured every week over a six week period as part of an agricultural expirement. Let y_{ij} be the diameter measurement from the ith apple in the jth week. Random coefficient models M_0, M_1 and M_2 were all fitted in R, and the following model comparisons were carried out:

CONTINUED OVERLEAF/

```
> anova(m0,m1)
Data: apple
Models:
m1: Diam ~ Time + (1 | AppleID) + (0 + Time | AppleID)
m0: Diam ~ Time + (Time | AppleID)
   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
m1  5 121.37 132.75 -55.684   111.37
m0  6 121.78 135.44 -54.887   109.78 1.5934      1     0.2068
> anova(m1,m2)
Data: apple
Models:
m2: Diam ~ Time + (1 | AppleID)
m1: Diam ~ Time + (1 | AppleID) + (0 + Time | AppleID)
   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
m2  4 137.91 147.02 -64.957   129.91
m1  5 121.37 132.75 -55.684   111.37 18.546      1  1.659e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ?’ 1
```

i. Based on these results, which of the three models would you select, and why?

ii. Which assumption is violated when comparing models M_1 and M_2? Why is this not a problem in this particular case?

[6 MARKS]

(e) Output from fitting model M_1 is shown below.

```
> summary(m1)
Linear mixed model fit by REML ['lmerMod']
Formula: Diam ~ Time + (1 | AppleID) + (0 + Time | AppleID)
   Data: apple

REML criterion at convergence: 117.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2206 -0.0257  0.0354  0.0907  2.5608

Random effects:
 Groups    Name        Variance Std.Dev.
 AppleID   (Intercept) 0.03776  0.1943
 AppleID.1 Time        0.03028  0.1740
 Residual              0.16775  0.4096
Number of obs: 72, groups:  AppleID, 12
```

CONTINUED OVERLEAF/

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.82772    0.12354  22.889
Time        -0.04800    0.05764  -0.833

Correlation of Fixed Effects:
     (Intr)
Time -0.393

> ranef(m1)
$AppleID
    (Intercept)        Time
1    0.025454395  0.06123113
4    0.018182339  0.07151399
5   -0.007426805  0.04424870
10   0.006071866  0.06182190
11  -0.010692557  0.04920600
13   0.033039132  0.07698327
14  -0.269180731 -0.47939795
15   0.063014383  0.10350599
17   0.149847413 -0.16014983
18  -0.006403647  0.04993039
19   0.003546664  0.06641095
25  -0.005452450  0.05469548
```

i. Predict the diameter of a new, unobserved apple on this tree in week 4.

ii. Predict the diameter of the apple with AppleID = 10 after 3 weeks.

[4 MARKS]

4. Researchers carried out a study to explore the effects of a new drug on arthritis severity. A group of 527 patients with arthritis were recruited, and the severity of their joint pain was observed over a six month period. To simplify the analysis, the severity score was dichotomised into two categories, "no/minimal joint pain" and "moderate/severe joint pain".

   This outcome variable, **pain**, was coded as 0 for "no/minimal pain" and 1 for "moderate/severe pain". Each patient had their baseline pain level observed and were randomly assigned either to a placebo or treatment group. The patients were then

CONTINUED OVERLEAF/

monitored once a month for 6 months.

The variables **id** (patient id number), **treat** (treatment status; 0 for placebo, 1 for drug) and **month** (month number; 0 to 6, with 0 indicating baseline values) were recorded.

The following SAS code was used to fit a GEE model to the data:

```
proc genmod data=arthritisstudy;
class pain id treat (param=ref ref='0');
model pain = treat month treat*month / dist=bin type3;
repeated subject = id/ corrw modelse type= ar(1) printmle;
run;
```

Selected output from the procedure is shown below:

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr> ChiSq
Intercept	1	3.7186	0.4351	2.8658	4.5714	73.04	<.0001
treat 1	1	-0.4091	0.4783	-1.3466	0.5284	0.73	0.3929
month	1	-1.1208	0.2283	-1.5683	-0.6733	24.10	<.0001
month*treat 1	1	-0.4113	0.2581	-0.9172	0.0946	2.54	0.1110
Scale	0	1.0000	0.0000	1.0000	1.0000		

Working Correlation Matrix

	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.3581	0.1282	0.0459	0.0164
Row2	0.3581	1.0000	0.3581	0.1282	0.0459
Row3	0.1282	0.3581	1.0000	0.3581	0.1282
Row4	0.0459	0.1282	0.3581	1.0000	0.3581
Row5	0.0164	0.0459	0.1282	0.3581	1.0000

GEE Fit Criteria	
QIC	1372.4128
QICu	1370.3721

CONTINUED OVERLEAF/

Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr> |Z|
Intercept	3.7201	0.4805	2.7783	4.6619	7.74	<.0001
treat 1	-0.3912	0.5233	-1.4169	0.6345	-0.75	0.4547
month	-1.1289	0.2497	-1.6183	-0.6395	-4.52	<.0001
month*treat 1	-0.4357	0.2671	-0.9592	0.0878	-1.63	0.1028

Analysis Of GEE Parameter Estimates Model-Based Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr> |Z|
Intercept	3.7201	0.5063	2.7278	4.7124	7.35	<.0001
treat 1	-0.3912	0.5489	-1.4670	0.6846	-0.71	0.4760
month	-1.1289	0.2415	-1.6022	-0.6556	-4.67	<.0001
month*treat 1	-0.4357	0.2673	-0.9596	0.0882	-1.63	0.1031
Scale	1.0000	.	.	.	.	.

Note: The scale parameter was held fixed.

Score Statistics For Joint Tests For GEE			
Source	DF	Chi-Square	Pr > ChiSq
treat	1	0.77	0.3795
month	1	19.87	<.0001
month*treat	1	2.17	0.1406

(a) The outcome variable contained 1701 missing values (out of a total of 527×7=3689). What assumption must we make about this missing data for our GEE analysis to be valid? Make sure you explain clearly what this means about the relationship between the missing and observed data. **[3 MARKS]**

(b) In the model, what indicates that a person who is designated as having moderate/severe joint pain is likely to continue to remain in that designation? **[2 MARKS]**

CONTINUED OVERLEAF/

(c) The researchers thought that there might not be a time decaying structure to the correlation and decided to fit an exchangeable GEE model to the data. The fit statistics for the exchangeable model are given below. Which model would you choose and why?

GEE Fit Criteria	
QIC	1372.8352
QICu	1370.4420

[2 MARKS]

(d) Does it seem like the probability of having moderate/severe pain over time is different for different treatment groups? **[2 MARKS]**

(e) Partial output from the model without interaction term is given below. Interpret the **month** and **treat** coefficient estimates in terms of odds of having moderate/severe pain.

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr> |Z|
Intercept	4.4522	0.2797	3.9040	5.0004	15.92	<.0001
treat 1	-1.2481	0.2271	-1.6932	-0.8030	-5.50	<.0001
month	-1.4965	0.0928	-1.6784	-1.3146	-16.13	<.0001

[4 MARKS]

(f) Using the output in part (e), what proportion of patients assigned to the drug are expected to experience moderate/severe pain in month 3? **[2 MARKS]**

(g) Give the name of an alternative model to GEE discussed in this course for analysing repeated data with non-normal distributions. How does it differ from GEE? In what situations might we prefer one or the other? **[5 MARKS]**

END OF QUESTION PAPER.