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. (a) Let Y_{ijk} be the cake loss obtained from the kth cake cooked by the jth chef using the ith recipe, with $i = 1, 2$, $j = 1, \ldots, 4$ and $k = 1, \ldots, 5$. **[2 MARKS]**

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 recipe, $\sum_{i=1}^{2} \alpha = 0$
- b_j is the random effect for the jth chef, $b_j \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$
- $(\alpha b)_{ij}$ is the random effect for the interaction between chef and recipe, $(\alpha b)_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_{AB}^2)$
- e_{ijk} are random errors, $e_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma_E^2)$

CONTINUED OVERLEAF/

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

(b) The full table with mean square errors is shown below

Source	DF	SS	MS
recipe (A)	1	5831	$MSA = \frac{5831}{1} = 5831$
chef (B)	3	434.58	$MSA = \frac{434.58}{3} = 144.86$
recipe*chef (AB)	3	401.29	$MSAB = \frac{401.29}{3} = 133.763$
error	32	1032	$MSE = \frac{1032}{32} = 32.25$

We wish to test the hypothesis: $H_A : \alpha_1 = \alpha_2$.

The test statistic is $F_A = \dfrac{MSA}{MSAB} = 43.6$ which is greater than the critical value $F(1, 3; 0.95) = 10.13$ so we reject the null hypothesis that $\alpha_1 = \alpha_2$ and conclude that there is a significant difference in cake loss between the two recipes.

[5 MARKS]

(c) We use the mean squares obtained in the previous part to obtain estimates for σ_E^2, σ_{AB}^2 and σ_B^2 as follows:

$$E(MSE) = \sigma_E^2 \Rightarrow \hat{\sigma}_E^2 = MSE = 32.25$$

$$E(MSAB) = \sigma_E^2 + 5\sigma_{AB}^2 \Rightarrow \hat{\sigma}_{AB}^2 = \frac{MSAB - \hat{\sigma}_E^2}{5} = \frac{133.763 - 32.25}{5} = 20.303$$

$$E(MSB) = \sigma_E^2 + 5\sigma_{AB}^2 + 10\sigma_B^2$$
$$\Rightarrow \hat{\sigma}_B^2 = \frac{MSB - 5\hat{\sigma}_{AB}^2 - \hat{\sigma}_E^2}{10} = \frac{MSB - MSAB}{10} = 1.110$$

[4 MARKS]

(d) Let $(\alpha b)_{i\cdot} = \frac{1}{4}\sum_{j=1}^{4}(\alpha b)_{ij}$ and $e_{i\cdot\cdot} = \frac{1}{20}\sum_{j=1}^{4}\sum_{k=1}^{3} e_{ijk}$

CONTINUED OVERLEAF/

$$\begin{aligned}\mathrm{Var}\big(\hat{\delta}\big) &= \mathrm{Var}\big(\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}\big) = \mathrm{Var}\big([(\alpha b)_{1\cdot} - (\alpha b)_{2\cdot}] + [e_{1\cdot\cdot} - e_{2\cdot\cdot}]\big) \\ &= \mathrm{Var}\big((\alpha b)_{1\cdot}\big) + \mathrm{Var}\big((\alpha b)_{2\cdot}\big) + \mathrm{Var}\big(e_{1\cdot\cdot}\big) + \mathrm{Var}\big(e_{2\cdot\cdot}\big) \text{ by independence} \\ &= \frac{\sigma^2_{AB}}{4} + \frac{\sigma^2_{AB}}{4} + \frac{\sigma^2_E}{20} + \frac{\sigma^2_E}{20} \\ &= \frac{5\sigma^2_{AB} + \sigma^2_E}{10}\end{aligned}$$

[3 MARKS]

(e) This is an example of a **crossed** design, because each possible combination of factor A and factor B is present - that is, every chef cooks both recipes.

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 chef cooked two separate recipes of their own, such that we had a total of 8 recipes, with no recipe being cooked by more than one chef.

[2 MARKS]

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

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

- μ is the overall mean
- γ is the fixed slope for baseline 5km time.
- x_{ik} is the baseline 5km time for the kth athlete using supplement i, with $k = 1, \ldots, 25$.
- α_i is the fixed effect for supplement type with $i = 1, 2$.
- β_j is the fixed effect for time in weeks with $j = 1, 2, 3, 4$.
- $(\alpha\beta)_{ij}$ is the interaction between supplement type and time.

[5 MARKS]

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

[2 MARKS]

CONTINUED OVERLEAF/

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.

[2 MARKS]

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

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

where ρ is the correlation between measurements taken 1 week apart and measurements taken k weeks apart have correlation ρ^k **[2 MARKS]**

iii. 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. **[2 MARKS]**

(c) Unstructured has the lowest AIC, while AR(1) gives the lower BIC. Both covariance structures allow the correlation between measurements from the same athlete to depend on the time distance between measurements.

AR(1) is a special case of the unstructured covariance structure, so a likelihood ratio test can be used to compare the two:

$$784.2 - 762.1 = 22.1 > \chi^2(8; 0.95) = 15.5.$$

The unstructured model is preferred.

[4 MARKS]

CONTINUED OVERLEAF/

(d) i. $\theta = 0$ and $\lambda = 0$, since both are the means of standard random effect terms.

ii. $\mathbf{V} = \mathbf{ZGZ}^\intercal + \mathbf{R}$

[3 MARKS]

3. (a) i. These random terms have the following distributional assumptions:

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

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

ii.

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

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

[4 MARKS]

(b) i. M_1: $\rho = 0$

ii. M_2: $\rho = 0$, $\sigma_1^2 = 0$

iii. M_3: $\rho = 0$, $\sigma_1^2 = 0$, $\sigma_0^2 = 0$

[3 MARKS]

(c) The student should provide three sketches which display an understanding of the models. Sketch 1 should have a number of regression lines with different slopes and intercepts. Sketch 2 should have a set of parallel regression lines with different intercepts. Sketch 3 should just have a single regression line (or several lines stacked on top of each other). See Figure 1 for examples.

[3 MARKS]

(d) i. The comparison of models M_0 and M_1 using the chi-squared test has a large p-value, indicating that the correlation between b_{0i} and b_{1i} is not significant. Therefore model M_1 is preferred.

CONTINUED OVERLEAF/

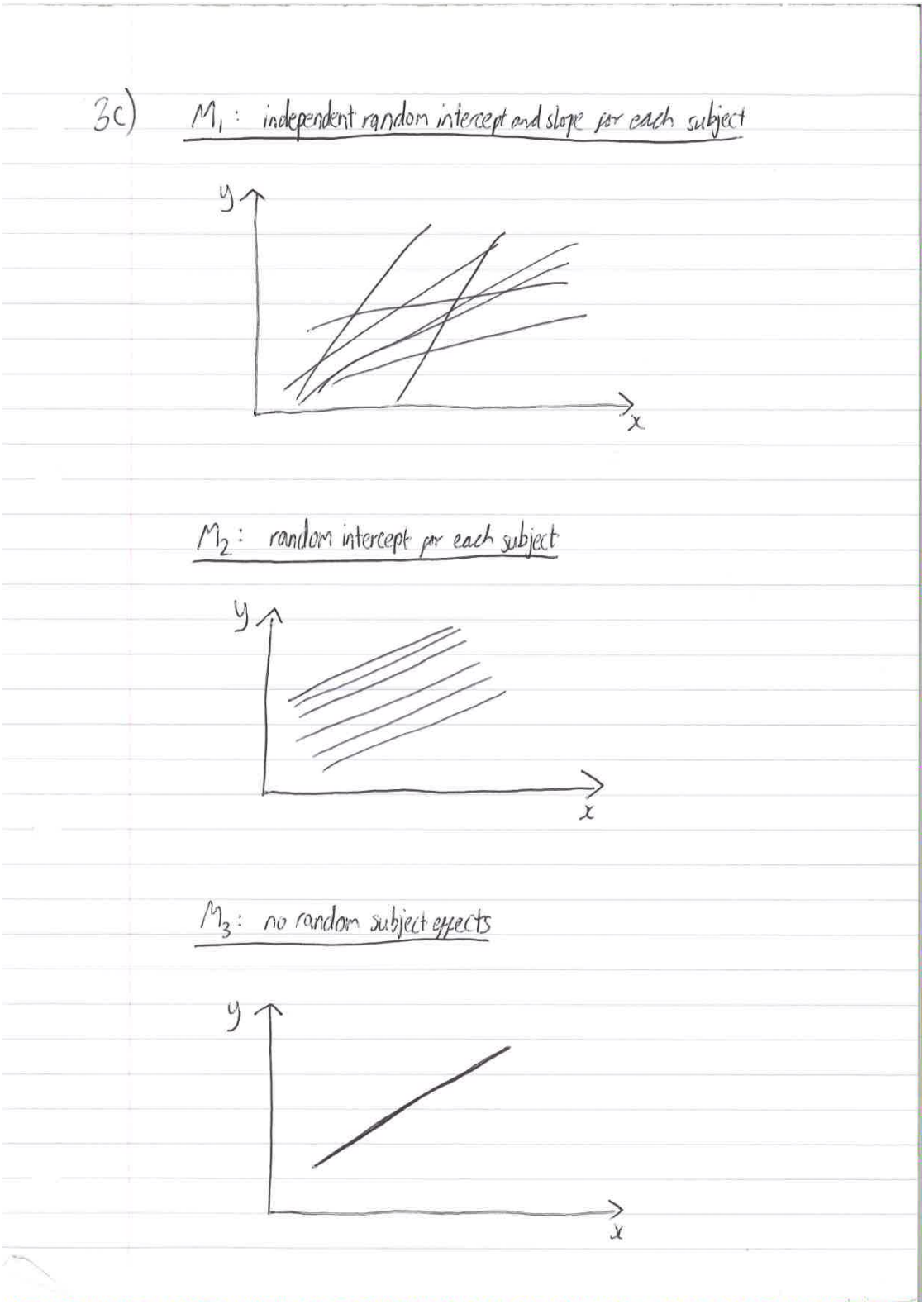

Figure 1: Sketches for Q3(c).

CONTINUED OVERLEAF/

The comparison of models M_1 and M_2 has a very small p-value, indicating that the random slope effect is needed in the model. So Model M_1 is preferred.

ii. The variance parameter being tested is on the boundary of possible values ($\sigma_1^2 = 0$), which violates the assumptions for chi-squared approximation. The p-value computed using the chi-squared approximation is generally larger than the p-value that would be obtained through simulation. In the model comparison for models M_1 and M_2 the chi-squared approximation gives a very small p-value, which implies that the 'true' p-value would be even smaller, and so the choice of model is the same whether the approximation is used or not.

[6 MARKS]

(e) i. The best prediction for an unobserved apple is given by the fixed effect terms in the model. This is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_{ij} = 2.82772 + (-0.04800) \times 4 = 2.636.$$

ii. The best prediction for the ith apple in the study is

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

For $i = 10$ and $x_{ij} = 3$ this equals

$$2.82772 + (-0.04800) \times 3 + 0.006071866 + 0.06182190 \times 3 = 2.875$$

[4 MARKS]

4. (a) The data must be MCAR, missing completely at random. This means the probability of an observation being missing must not depend on either the observed or missing values.

[3 MARKS]

(b) The working correlation matrix shows positive correlations from month to month for the same subject. These range between 0.0164 (weak) and 0.3581 (moderately strong).

[2 MARKS]

(c) The fit statistics for the AR(1) model are slightly smaller than those for the exchangable GEE. Therefore we might prefer the AR(1) model. However, the difference in fit statistics between these models is very small.

[2 MARKS]

(d) The term for the interaction between month and treatment in our model has a high p-value (much larger than 0.05), therefore it does not appear that this term

CONTINUED OVERLEAF/

is necessary in the model. There does not appear to be a significant difference between the treatment group and the placebo group in terms of the change in probability of moderate/severe pain over time.

[2 MARKS]

(e) The odds of having moderate/severe pain get multiplied by a factor of exp(-1.4965) = 0.2239 for each additional month. The longer the patient is in the study, the less likely they are to have moderate/severe pain.

Patients who are assigned to the drug (**treat**=1) have lower odds of having moderate/severe pain by a factor of exp(-1.2481) = 0.2870.

[4 MARKS]

(f) The probability of moderate/severe pain (**pain**=1) for a patient who is assigned to the drug (**treat**=1) in month 3 (**month**=3) is given by:

$$\frac{\exp(4.4522 - 1.2481 - 1.4965 \times 3)}{1 + \exp(4.4522 - 1.2481 - 1.4965 \times 3)} = 0.2166$$

[2 MARKS]

(g) An alternative model would be the generalised linear mixed model (GLMM).

In the GLMM, the linear predictor can contain random effects which are normally distributed. The GEE models the mean response and covariance separately and incorporates the dependence between within subject observations in the correlation matrix structure.

We use GEEs when the main structure of interest is the population level mean relationship and the correlation of repeated measurements is a nuisance. GLMMs are used when we want an idea of how the mean relationship varies across subjects or when we might want predictions for particular subjects.

[5 MARKS]

END OF QUESTION PAPER.