MAS474 Extended linear models
2016-17 Exam Solutions

1. (i.) (routine)

$$Y_{ijk} = \begin{cases} \mu_b + b_i + b_{ij} + \epsilon_{ijk} \text{ if boy} \\ \mu_g + b_i + b_{ij} + \epsilon_{ijk} \text{ if girl} \end{cases} \checkmark\checkmark$$

where $b_i \sim N(0, \sigma_1^2)$, $b_{ij} \sim N(0, \sigma_2^2)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$ $\checkmark$

(ii.) (routine) Interest lies in the difference between boys and girls, and so these must be fixed effects. $\checkmark$

We need to account for differences between schools and classes within schools (which may be gender segregated) but are not directly interested in these differences, and so these are modelled as random effects. $\checkmark$

Classes are nested within schools and are not comparable between schools, which is why we have a b_{ij} term not a b_j term in the model $\checkmark$

(iii.) (routine)

$$\mu_b = 39.571 \qquad \mu_g = 44.588 \checkmark$$

$$\sigma_1^2 = 53.93 \qquad \sigma_2^2 = 34.15 \qquad \sigma^2 = 382.99 \checkmark$$

(iv.) (unseen)

$$\begin{aligned} Cov(Y_{ijk}, Y_{ij'k'}) &= Cov(b_i + b_{ij} + \epsilon_{ijk}, b_i + b_{ij'} + \epsilon_{ij'k'}) \\ &= Var(b_i) = \sigma_1^2 \checkmark\checkmark \end{aligned}$$

and

$$Var(Y_{ijk}) = \sigma_1^2 + \sigma_2^2 + \sigma^2 \checkmark$$

Hence

$$Cor(Y_{ijk}, Y_{ij'k'}) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + \sigma^2} = \frac{34.14}{34.15 + 53.93 + 382.99} = 0.0725 \checkmark$$

(v.) (routine) These are to check that the random effects do indeed have a normal distribution. $\checkmark$

They look fine, as the points appear to lie on a straight line as expected if the normal assumption is true. $\checkmark$

(vi.) (routine) Bootstrap hypothesis test $\checkmark$

Testing $H_0 : \mu_b = \mu_g$ vs $H_1 : \mu_b \neq \mu_g$ i.e. no difference between boys and girls $\checkmark$

Estimated p-value is 5×10^{-4}, so strong evidence to reject H_0 in favour of H_1, i.e., there is a difference between boys and girls. $\checkmark\checkmark$

(vii.) (unseen) Something like

```
lmer(english ~ gender +year+(1|school/class/id))
```

or variations thereof. ✓✓

2. (i.) (unseen)

$$\begin{aligned} L(\lambda|Y) &= \prod_{i:Y_i=1} \mathbb{P}(X_i \le z_i) \prod_{i:Y_i\neq 1} \mathbb{P}(X_i > z_i) \checkmark \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \le z_i)^{Y_i} \mathbb{P}(X_i > z_i)^{1-Y_i} \checkmark \\ &= \prod_{i=1}^{n} (1 - e^{-\lambda z_i})^{Y_i} (e^{-\lambda z_i})^{1-Y_i} \checkmark \end{aligned}$$

It would be hard to directly maximize this wrt λ, hence we might use the EM algorithm. ✓

(ii.) (unseen) The easiest way is to note that the distribution of $X|X > z_i$ is still exponential(λ) by the memoryless property (✓), and so

$$\mathbb{E}(X|X > z_i) = z_i + \frac{1}{\lambda} \checkmark\checkmark$$

Some students may attempt to do this by deriving the pdf and then calculating the integral, which is fine as well, but takes more work.

(iii.) (unseen)

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}(\mathbb{E}(X|Y)) \\ &= \mathbb{E}(X|Y=0)\mathbb{P}(Y=0) + \mathbb{E}(X|Y=1)\mathbb{P}(Y=1) \checkmark \\ &= \mathbb{E}(X|X > z_i)\mathbb{P}(X > z_i) + \mathbb{E}(X|X \le z_i)\mathbb{P}(X \le z_i) \checkmark \\ \frac{1}{\lambda} &= (z_i + \frac{1}{\lambda})e^{-\lambda z_i} + \mathbb{E}(X_i|X_i \le z_i)(1 - e^{-\lambda z_i}) \checkmark\checkmark \end{aligned}$$

Thus

$$\mathbb{E}(X_i|X_i \le z_i) = \frac{1 - (1 + \lambda z_i)e^{-\lambda z_i}}{\lambda(1 - e^{-\lambda z_i})} \checkmark\checkmark$$

(iv.) (unseen) Introduce X as the missing data. Then

$$\begin{aligned} L(\lambda|X, y, z) &= \prod \lambda e^{-\lambda X_i} \mathbb{I}_{Y_i = \mathbb{I}_{X_i \le z_i}} \\ &= \lambda^n e^{-\lambda \sum X_i} \mathbb{I}_{Y_i = \mathbb{I}_{X_i \le z_i} \forall i} \checkmark \end{aligned}$$

Thus

$$\begin{aligned} Q(\lambda, \lambda^{(m)}) &= \mathbb{E}_{X|\lambda^{(m)},y,z} \left(\log L(\lambda|X, y, z)\right) \checkmark \\ &= n \log \lambda - \lambda \sum \mathbb{E}_{X|\lambda^{(m)},y,z} X_i + \mathbb{E}(\log \mathbb{I}_{Y_i = \mathbb{I}_{X_i \le z_i} \forall i} \mid X_i \le z_i) \checkmark \\ &= n \log \lambda - \lambda \sum_{i=1}^{n} L_i^{(m)Y_i} R_i^{(m)1-Y_i} + 0 \checkmark \end{aligned}$$

where $L_i^{(m)}$ and $R_i^{(m)}$ are estimates of $\mathbb{E}(X_i|X_i \leq z_i)$ and $\mathbb{E}(X_i|X_i > z_i)$ calculated using $\lambda^{(m)}$.✓The expectation of the log of the indicator function is 0 as we are conditioning upon the event $\{X_i \leq z_i\}$, and so the indicator function always takes value 1, and thus the log is zero.

This is minimized at

$$\hat{\lambda} = \lambda^{(m+1)} = \frac{n}{\sum_{i=1}^{n} L_i^{(m)Y_i} R_i^{(m)1-Y_i}}.$$✓

The EM algorithm then iterates from some starting value of $\lambda^{(0)}$✓

- Calculate $L_i^{(m)}, R_i^{(m)}$ given $\lambda^{(m)}$ ✓
- Calculate $\lambda^{(m+1)}$ given $L_i^{(m)}, R_i^{(m)}$ ✓

(maximum of 7 ✓)

3. (i.) (routine)

   (a) MAR ✓

   (b) MCAR ✓

   (c) NMAR ✓

   a and b are ignorable ✓

   (ii.) (routine)

   (a) $\frac{-9.5+6.4}{2} = -1.55$ ✓

   (b) $\frac{-9.5+6.4-13.7}{3} = -5.6$ ✓

   (c) Same as for available-case: -5.6 ✓

   (d)

$$\hat{y}_4 = 2.866667 + 1.962963 \times 16 = 34.27407 ✓$$

   and thus

$$\frac{-9.5 + 6.4 - 13.7 + 34.274}{4} = 4.3685 ✓$$

   is the regression imputation estimate.

   (iii.) (unseen)

   (a) Multiple imputation using chained equations ✓

   It creates $m = 5$ ✓complete datasets by replacing the missing values using a variety of imputation methods.✓The approach begins by filling in missing values by sampling with replacement ✓. It then

   - replaces missing values in AttDebt using linear regression given the other values
   - replaces missing values in CCard using multinomial regression given the other values
   - replaces missing values in BuildSoc using logistic regression given the other values
   - replaces missing values in LOC using linear regression regression given the other values ✓✓

   It uses stochastic imputation ✓meaning that a random error is added and parameter uncertainty is accounted for ✓(Bayesian regression approach).
   It then iterates through the 4 steps above until convergence ✓.
   (maximum of 6 ✓)

   (b) (unseen) We can estimate the expected value by combing the parameter estimates from each imputed datasets

$$\begin{aligned}\mathbb{E}(\beta_3|Y_{obs}) &= \frac{1}{m}\sum \hat{\beta}_3^{(i)} ✓ \\ &= \frac{-1.12 - 2.11 - 0.53 - 1.32 - 1.01}{5} = -1.22\end{aligned}$$

To calculate the variance, we need to use the corrected version of the posterior variance

$$Var(\theta|Y_{obs}) \approx \bar{V} + (1 + \frac{1}{m})B$$

where

$$\bar{V} = \frac{1}{5}(0.67 + 0.78 + 0.46 + 0.90 + 0.46) = 0.654$$

is the average within imputation variability.

and

$$B = \frac{5}{4}(\frac{1}{5}\sum \beta_3^{(i)2} - \bar{\beta}_3^2) = 0.33307$$

is the between imputation variability. Alternatively, this can be read from the R output.

Thus

$$Var(\theta|Y_{obs}) \approx 0.654 + (1 + 1/5)0.333 = 1.0536$$

and so the estimated standard error is $\sqrt{1.0536} = 1.0264$.