MAS473 Extended linear models
2014-15 Exam Solutions

(1)(i) Model is

$$Y_{ijk} = \mu + \tau_i + b_{ij} + \epsilon_{ijk},$$

[1 mark]

where

Y_{ijk} is the blood pressure for drug i, patient j within group i, and occasion k,

$i = 1, 2$, $j = 1, \ldots, 20$ and $k = 1, \ldots, 5$

[1 mark for correct subscript for random effect, 1 mark for everything else]

$b_{ij} \sim N(0, \sigma_1^2)$

$\epsilon_{ijk} \sim N(0, \sigma^2)$.

$\tau_1 = 0$.

[1 mark]

(1)(ii) Patient has been modelled as a random effect, as the interest is likely to be in the effect of the drug on the population of patients, and not just the forty patients in the study.

[1 mark]

(1)(iii)

$$\begin{aligned} \hat{\mu} &= 140.8649, \\ \hat{\tau}_2 &= -1.9624, \\ \hat{\sigma}_1^2 &= 10.575, \\ \hat{\sigma}^2 &= 3.753. \end{aligned}$$

[1 mark]

(1)(iv)

$$\widehat{Var}(Y_{ijk}) = \widehat{Var}(b_{ij}) + \widehat{Var}(\epsilon_{ijk}) = 14.328.$$

[M1 A1]

(1)(v) The purpose of the command is to test the assumption that $b_{ij} \sim N(0, \sigma_1^2)$

[1 mark]

The reference line should have gradient 3.252.

[1 mark]

(1)(vi)

$$
\begin{aligned}
Var(\bar{Y}_{2..} - \bar{Y}_{1..}) &= Var(\bar{Y}_{2..}) + Var(\bar{Y}_{1..}) \\
&= 2Var(\bar{Y}_{1..}) \\
&= 2Var\left(\frac{1}{100}\sum_{j=1}^{20}\sum_{k=1}^{5}(b_{1j} + \epsilon_{1jk})\right) \\
&= \frac{2}{100^2}Var\left(5\sum_{j=1}^{20} b_{1j} + \sum_{j=1}^{20}\sum_{k=1}^{5}\epsilon_{1jk}\right) \\
&= \frac{2\times 25}{100^2} \times 20Var(b_{1j}) + \frac{2}{100}Var(\epsilon_{1jk})
\end{aligned}
$$

[3 marks]

Estimated standard error is

$$\sqrt{\frac{2\times 25}{100^2}\times 20 \times 10.575 + \frac{2}{100}\times 3.753} = 1.064.$$

[1 mark]

(1)(vii) The procedure that has been used is a parametric bootstrap hypothesis test

[1 mark]

The null hypothesis is that $\tau_2 = 0$.

[1 mark]

The output gives an estimated p-value of 0.12, suggesting no evidence against the null hypothesis: there is no evidence that the drug works better than placebo.

[1 mark]

(1)(viii) The t-test has found a significant difference between the drug and placebo groups, unlike the bootstrap hypothesis test

[1 mark]

However, the t-test assumes the observations are independent, but this assumption is not valid, as each patient's pressure is observed 5 times, and observations on the same patient are correlated.

[1 marks]

This gives a denominator in the t statistic that is too small, which has made the test statistic larger and given the significant result.

[1 mark]

(2)(i) $E(Y_i) = \frac{exp(\eta_i)}{1+exp(\eta_i)}$ for logit **[1 mark]**, $E(Y_i) = \Phi(\eta_i)$ for probit.**[1 mark]**

(2)(ii) Comparing model 1 and 3. $H_0 : \beta_2 = 0$. $\Delta D = 20.78$ on 1 df so H_0 rejected. **[2 marks]**
Comparing model 2 and 3. $H_0 : \beta_1 = 0$. $\Delta D = 41.83$ on 1 df so H_0 rejected. **[2 marks]**

Comparing model 3 and 4. $H_0 : \beta_3 = 0$. $\Delta D = 6.32$ on 1 df so H_0 rejected. **[1 mark]** Interaction model (model 4) looks best. Log odds of lung cancer depends on both age and smoking status but not additively . **[1 mark]**

(2)(iii) $\chi^2_{76,0.95} = 97.35$. **[1 mark]** Observed residual deviance of 32.26 is consistent with a χ^2_{76} so model looks a reasonable fit. **[1 mark]**

(2)(iv) $\eta = -121.456 + 118.886 + 70 * (1.854 - 1.777) = 2.82$**[1 mark]**
odds=exp(2.82)=16.8**[1 mark]**

(2)(v) Let x be the age at the which the estimated odds coincide. Then

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_2 x &= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 x \textbf{[1 mark]}\\ x &= -\frac{\hat{\beta}_1}{\hat{\beta}_3} = -118.886/-1.777 = 66.9 \text{ years}\textbf{[1 mark]}\end{aligned}$$

(2)(vi)

$$\begin{aligned}L &= \prod_{\text{cases}} P(Y_i = 1; \eta_i) \prod_{\text{controls}} P(Y_i = 0; \eta_i)\textbf{[1 mark]}\\ &= \prod_{\text{cases}} \left(\frac{exp(\eta_i)}{1 + exp(\eta_i)}\right) \prod_{\text{controls}} \left(\frac{1}{1 + exp(\eta_i)}\right) \textbf{[1 mark]}\\ l &= \sum_{\text{cases}} (\eta_i - log(1 + exp(\eta_i))) - \sum_{\text{controls}} (log(1 + exp(\eta_i)))\\ &= \sum_{\text{cases}} (\beta_0 + \beta_1 x_{1i}) - \sum_{\text{all}} log(1 + exp(\beta_0 + \beta_1 x_{1i}))\textbf{[1 mark]}\\ \frac{\partial l}{\partial \beta_0} &= \sum_{\text{cases}} 1 - \sum_{\text{all}} \left(\frac{exp(\beta_0 + \beta_1 x_{1i})}{1 + exp(\beta_0 + \beta_1 x_{1i})}\right) \textbf{[1 mark]}\\ &= \sum_{\text{cases}} 1 - \sum_{\text{all}} \left(1 - \frac{1}{1 + exp(\beta_0 + \beta_1 x_{1i})}\right)\\ \frac{\partial^2 l}{\partial {\beta_0}^2} &= -\sum_{\text{all}} \left(\frac{exp(\beta_0 + \beta_1 x_{1i})}{1 + exp(\beta_0 + \beta_1 x_{1i})}\right) = -\sum_{\text{all}} \left(\frac{exp(\eta_i)}{1 + exp(\eta_i)}\right) \textbf{[1 mark]}\end{aligned}$$

It is used in the information matrix and so gives information on the parameter standard errors. **[1 mark]**

(3)(i) Percentages to the nearest whole number are **[1 mark]**
There is some difference in the proportion objecting by ownership **[1 mark]** but the difference by birthplace is more pronounced **[1 mark]**.
Or 2 other sensible observations.

(3)(ii) Comparing model 1 and 2: $\Delta D = 10.46$ on 2 df so H_0 that additional parameters are zero is rejected. **[2 marks]**
Comparing model 1 and 3: $\Delta D = 11.38$ on 1 df so H_0 that additional parameters are

local				incomer			
owner	object	approve		owner	object	approve	
multinational	54			multinational	56		
local_company	16			local_company	56		
community	33			community	53		

zero is rejected. **[2 marks]**
Model 3 is the best model since it has more df than model 2 but has a lower χ^2 value. **[1 mark]**

(3)(iii) $\chi^2_{4,0.95} = 9.49$ and since the residual deviance is 15.6 is is not a particularly good fit. **[1 mark]**
Yes it does agree with the observations from part (i) since the chosen model represents homogeneity with respect to ownership but not birthplace. **[1 mark]**

(3)(iv) It says that the probability of objecting is constant (1/2) and doesn't depend on ownership or birthplace. **[1 mark]**

(3)(v) There are 80 objections and 111 approvals so the fitted value is $6 \times 80/191 = 2.51$ **[M1 A1]**. The Pearson residual is therefore $\frac{2-2.51}{\sqrt{2.51}} = -0.32$ **[1 for numerator(FT wrong fitted), 1 for correct answer (FT wrong fitted)**.

(3)(vi)

$$\begin{aligned}
y_{ijk} &\sim Po(\mu_{ijk}) \sim Po(n_{jk}\pi_{ijk}) \sim Po(n_{jk}\pi_{ik})\textbf{[1 mark]} \\
L &= \prod_i \prod_j \prod_k \left\{ \frac{\exp(-\mu_{ijk})[\mu_{ijk}]^{y_{ijk}}}{y_{ijk}!} \right\} \textbf{[1 mark for indep Poissons]} \\
l &= \sum_i \sum_j \sum_k \{-n_{jk}\pi_{ik} + y_{ijk}[log n_{jk} + log\pi_{ik}] - log(y_{ijk}!)\} \textbf{ [1 mark]} \\
\frac{\partial l}{\partial \pi_{ik}} &= \sum_j \left\{ -n_{jk} + \frac{y_{ijk}}{\pi_{ik}} \right\} \textbf{ [1 mark]} \\
\widehat{\pi}_{ik} &= \frac{\sum_j y_{ijk}}{\sum_j n_{jk}}\textbf{[1 mark]}
\end{aligned}$$