Solution proposal finals STK3100/4100-f16

Problem 1

a) The density can be written

$$\begin{aligned} f(y;\mu,\nu) &= \frac{y^{-1}}{\Gamma(\nu)}(\frac{y\nu}{\mu})^{\nu}\exp(-y\nu/\mu), y>0. \\ &= \frac{1}{y}\frac{(\nu y)^{\nu}}{\Gamma(\nu)}\exp(-y\nu/\mu-\nu\log(\mu)) \\ &= \frac{1}{y}\frac{(\nu y)^{\nu}}{\Gamma(\nu)}\exp(\frac{y(-\frac{1}{\mu})-\log(\mu)}{\frac{1}{\nu}}). \end{aligned}$$

from which we see that $\phi = 1/\nu$, $c(y;\phi) = \frac{1}{y}\frac{(\nu y)^{\nu}}{\Gamma(\nu)}$, $\theta = -1/\mu$ and $a(\theta) = \log(\mu) = \log(-1/\theta) = -\log(-\theta)$.

Since $a'(\theta) = -\frac{1}{\theta} = \mu$, $E(y) = \mu$.

b) The canonical link is obtained from $\theta = \eta$ where η is the predictor. The link is given by $\eta = g(\mu)$ so $-1/\theta = \mu = g^{-1}(\eta) = g^{-1}(\theta)$. Hence $g(-1/\theta) = \theta$ or $g(\theta) = -1/\theta$, i.e. the inverse. The problem with this link is that since $\mu > 0$, $\theta < 0$, the linear predictor will also be negative and more importantly not having the entire real line as range. This is not a good property, so the canonical link is not much used for gamma distributed response. Instead the log -link is much used.

Problem 2

a) The number of persons in each combination of the covariates is large. One can then think of the number of accidents as the sum of a large number of Bernoulli trials where the number of trials is large, and the success parameter, in this case the probability of being killed in a traffic accident, is small. The sum of the successes of Bernoulli trials has a Binomial distribution. For small success probabilities and large number of trials the probabilities in the Binomial distribution are close to the probabilities in a Poisson distribution. Hence it is reasonable to consider the responses as Poisson distributed in this case.

The number of groups is $2 \times 8 = 16$ and the number of parameters is $1+(2-1)+(8-1) = 9$ which means that the deviance is approximately χ^2-distributed with 16-9=7 degrees of freedom, cf. de Jong and Heller page 72. Then the probability for a value larger than the observed deviance is 0.19, so the fit is satisfactory.

b) The expected number of deaths in each group will depend on the size of the population. If $n_{ij}, i = 1,2,\ j = 1,\ldots,8$, are the population sizes, the expected number of deaths will be $n_{ij}f(gender_i, age_j)$.

Using the log link where $\eta = \log(\mu)$, or $\mu = \exp(\eta)$, the expected number of deaths in group ij will have the form $(n_{ij}/sc)\exp(\eta) = \exp(\log(n_{ij}/sc) + \eta)$. Remark that $exp(\eta)$ will have the interpretation as the rate pr sc units. The coefficient of $\log(n_{ij}/sc)$ is equal to one, which means that it must be specified as an offset.

c) The base group for gender is men and for age 0-17, and from the R-output one can see that the population is counted in 100000 individuals. Hence $\exp(\beta_0)$ is the rate of killed per 100000 in the base group, $(n_{11}/100000)\exp(\beta_0)$ is the expected number of death in this group, and $(n_{11}/100000)\exp(\hat{\beta}_0)$ is the fitted value for this combination of the factor levels.

The gender effect is estimated as $\hat{\beta}_1 = -1.0212$. The Wald statistic for the test $H_0 : \beta_1 = 1$ vs $H_1 : \beta_1 \neq 1$ is $\frac{(\hat{\beta}_1 - 1}{\hat{se}(\hat{\beta}_1)}$ where $se_{\hat{\beta}_1}(\beta_1)$ is the standard error of β_1 and $\hat{se}(\hat{\beta}_1) = se_{\hat{\beta}_1}(\hat{\beta}_1)$. From the output $se = 0.1858$, so the test statistic is $-0.0212/0.1858 = -0.1141$ and the p-value is $2P(Z > 0.1141) = 0.91$ where Z is a standard normally distributed variable, so there is no reason to reject the null hypothesis.

d) The estimated predictor for women of age 45-54 is $0.1506 - 1.0212 + 1.5366 = 0.6660$, so the estimated rate of deaths pr 10000 is $exp(0.6660) = 1.9465$. The population in this group is 3.38505×10000 so the fitted value is $3.38505\exp(0.6660) = 6.5888$ and residual is $2 - 6.5888 = -4.5888$ since the number of fatal accidents was 2.

e) The y be the vector of responses where the first 8 elements are the number of accidents for men in age group $i = 1, \ldots, 8$ and the 8 last ones are the number of accidents for women. The design matrix is then

$$X = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

The fitted values $\hat{\mu}$ satisfies the first order requirements $\frac{\partial l}{\partial \beta} = X'D(y - \hat{\mu}) = 0$ where l is the log likelihood function and $D = diag(\frac{\partial \theta_i}{\partial \eta_i})$. For the canonical link $\theta = \eta$ so $D = I_{16}$, the identity matrix of order 16. Then $X'y = X'\hat{\mu}$. The coefficient for gender is β_1 so $\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{16} x_{i2}(y_i - \hat{\mu}_i) = 0$. But $x_{i2} = 0,\ i = 1, \ldots, 8$ and $x_{i2} = 1,\ i = 9, \ldots, 16$. Hence $\sum_{i=9}^{16} y_i = \sum_{i=9}^{16} \hat{\mu}_i$. The left hand side is the sum of accidents among women and the right hand side is the sum of fitted values for women.

Problem 3

a) Define the matrices

$$X_i = \begin{pmatrix} 1 & redage_{i1} \\ 1 & redage_{i2} \\ 1 & redage_{i3} \\ 1 & redage_{i4} \end{pmatrix},\ i = 1, \ldots, 5.$$

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}y_{i4})' = (\text{bone}_{i1}, \text{bone}_{i2}, \text{bone}_{i3}, \text{bone}_{i4})'$ be the responses. Then the model may be written on matrix form as

$$\mathbf{y}_i = X_i \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + Z_i \begin{pmatrix} b_{i,1} \\ b_{i,2} \end{pmatrix} + \varepsilon_i$$

where $Z_i = X_i$ and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$.

Here X_i is the design matrix for the fixed effects part. The random vectors $\mathbf{b}_i = (b_{i,1}, b_{i,2})'$ define the random part. The fitted values are in this case 5 non-paralell lines (random slope and intercept). The model is appropriate when it is the distribution of the intercepts and slopes which is of primary interest, not the intercept and slope for particular units.

The assumptions are that the random vectors $\mathbf{b}_i = (b_{i,1}, b_{i,2})$ and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$ are independent and with multinormal distributions with expectation zero. The covariance matrix of $\mathbf{b}_i$ has the form $D = \begin{pmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{pmatrix}$. The covariance of ε_i, Σ_i can be general, but is often of the form $\sigma^2 I_4$ where I_4 is a 4×4 identity matrix.

b) Since $\mathbf{b}_i$ and ϵ_i are indrpendent multinormally distributed also the distribution of $\mathbf{y}_i$ is multinormal.

The expectation of the response is $X_i\beta$ where X_i are the design matrix where the elements are the values of the covariates in cluster i.

Using that $\mathbf{b_i}$ and ϵ_i are independent the covariance matrix of the response is $V_i = Cov(\mathbf{y}_i) = Cov(Z_i\mathbf{b}_i) + Cov(\epsilon_i)$. Since $Cov(Z_i\mathbf{b_i}) = Z_iCov(\mathbf{b}_i)Z_i' = Z_iDZ_i'$ and $Cov(\epsilon_i) = \Sigma_i$, the marginal covariance matrix is $V_i = Z_iDZ_i' + \Sigma_i$.

Referring to the R-output $X_i = Z_i$ contains the measured values of the centered age of the five boys at the four occasions, i.e.

$$Z_i = \begin{pmatrix} 1 & -0.75 \\ 1 & -0.25 \\ 1 & 0.25 \\ 1 & 0.75 \end{pmatrix},$$

$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)' = (52.690, 1.424)'$, $\hat{D} = \begin{pmatrix} 0.8172867^2 & 0.8172867 \times 0.732361 \\ 0.8172867 \times 0.7323611 \times 0.586 & 0.7323611^2 \end{pmatrix}$ and $\hat{\Sigma}_i = 0.2939400^2 I_4$ for $1 = 1, \cdots, 5$.

c) Since $\mathbf{y}_1, \ldots, \mathbf{y}_5$ are independent and only $\mathbf{y}_i$ is correlated with $\mathbf{b}_i$, $E[\mathbf{b}_i|\mathbf{y}_1, \ldots, \mathbf{y}_5] = E[\mathbf{b}_i|\mathbf{y}_i]$.

But $(\mathbf{b}_i, \mathbf{y}_i)'$ is multinormally distributed with expectation $(0, X_i\beta)'$ and covariance matrix

$$\begin{pmatrix} D & DZ_i' \\ Z_iD & Z_iDZ_i' + \Sigma_i \end{pmatrix}.$$

Hence

$$E[\mathbf{b}_i|\mathbf{y}_i] = E[\mathbf{b}_i] + Cov(\mathbf{b}_i, \mathbf{y}_i)[Var(\mathbf{y}_i)]^{-1}(\mathbf{y}_i - E[\mathbf{y}_i]) = DZ_i'(Z_iDZ_i' + \Sigma_i)^{-1}(\mathbf{y}_i - X_i\beta).$$

By plugging inn the REML estimates for D and Σ_i from part b) and the estimates for β from the R-output, i.e. $\hat{\beta}_0 = 52.690$ and $\hat{\beta}_1 = 1.424$, the random effects $\mathbf{b}_i$ can be estimated.