STATISTICAL MODELLING Part IIC.

Example Sheet 3 (of 4) RN/Lent 2010

1. Return to the `mammals` data from Practical 4. Let x_i denote the body weight of the ith mammal, and let Y_i denote its brain weight. The second model fitted assumed that $Y_1, \ldots, Y_n$ were independent, with

$$\log Y_i = \alpha + \beta \log x_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$. Use `confint` to find a 95% confidence interval for β, and check the calculation yourself. Find also an elliptical 95% confidence set for $(\alpha, \beta)^T$, explaining why `confint` is not appropriate here. Give a prediction $\hat{Y}$ of the brain weight Y^* of a new mammal with body weight 30kg, together with a 95% prediction interval. Is it the case that $\mathbb{E}(Y^*) = \mathbb{E}(\hat{Y})$?

2. (a) Let X and Y be independent random variables with densities $f_X(x)$ and $f_Y(y)$ respectively. Then the density $f_Z(z)$ of $Z = X+Y$ is the convolution of $f_X(x)$ and $f_Y(y)$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx.$$

Show that if $Y_1, \ldots, Y_n$ are independent and Y_i has density $f_{Y_i}(y_i)$, then $S = Y_1 + \ldots + Y_n$ has density

$$f_S(s) = \int_{\mathcal{S}} \prod_{i=1}^{n} f_{Y_i}(y_i)\, dy_1 \ldots dy_n,$$

where $\mathcal{S} = \{(y_1, \ldots, y_n) : y_1 + \ldots + y_n = s\}$.

(b) Let $Y_1, \ldots, Y_n$ be independent and identically distributed with density $f(y; \theta) = \exp\{y\theta - K(\theta)\} f_0(y)$ for $y \in \mathcal{Y} \subseteq \mathbb{R}$, $\theta \in \Theta \subseteq \mathbb{R}$. Show that S has density

$$f_S(s; \theta) = e^{\theta s - nK(\theta)} f_S(s), \quad s \in \mathcal{S},\ \theta \in \Theta,$$

where $f_S(s)$ is the density of S when $Y_1, \ldots, Y_n$ are independent with density $f_0(y)$, $y \in \mathcal{Y}$.

3. Let Y have a model function of exponential dispersion family form. Compute the cumulant generating function of Y and deduce expressions for the mean and variance of Y.

4. We say Y has the inverse Gaussian distribution with parameters ϕ and λ, and write $Y \sim IG(\phi, \lambda)$ if its density is

$$f_Y(y; \phi, \lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi} y^{3/2}} e^{\sqrt{\lambda\phi}} \exp\left\{-\frac{1}{2}\left(\frac{\lambda}{y} + \phi y\right)\right\},$$

$y \in (0, \infty)$, $\lambda \in (0, \infty)$, $\phi \in (0, \infty)$. Compute the cumulant generating function of Y, and find its mean and variance. By making a carefully chosen reparametrisation from (ϕ, λ) to (μ, σ^2), deduce that $Y \sim ED(\mu, \sigma^2 V(\mu))$, $\mu \in \mathcal{M}$, $\sigma^2 \in \Phi$, where $\mathcal{M}$, Φ and $V(\mu)$ should be specified, together with the canonical link function for this family.

5. Let $Y_1, \ldots, Y_n$ be independent random variables with

$$Y_i \sim ED\left(\mu, \frac{\sigma^2}{w_i} V(\mu)\right), \quad \mu \in \mathcal{M},\ \sigma^2 \in \Phi \subseteq (0, \infty),$$

where $w_1, \ldots, w_n$ are known constants. Let $w_+ = \sum w_i$. By considering cumulant generating functions, show that

$$\frac{1}{w_+} \sum_{i=1}^{n} w_i Y_i \sim ED\left(\mu, \frac{\sigma^2}{w_+} V(\mu)\right), \quad \mu \in \mathcal{M}.$$

Deduce the distribution of the sample mean of a random sample from

(a) $N(\mu, \sigma^2)$

(b) the gamma distribution with mean $\nu\phi$ and variance $\nu\phi^2$

(c) $IG(\phi, \lambda)$.

(d) Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim \frac{1}{n_i}\mathrm{Bin}(n_i, p)$ for $i = 1, \ldots, n$, and let $N = \sum n_i$. What is the distribution of $\frac{1}{N} \sum n_i Y_i$?

6. Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$, where $\mu_i = \alpha + \beta x_i$, and assume for simplicity that σ^2 is known. Show that only one iteration of the Fisher scoring method is required to attain the maximum likelihood estimator $(\hat{\alpha}, \hat{\beta})^T$, regardless of the initial values for the algorithm. What feature of the log-likelihood function ensures that this is the case?

7. Let Y have the exponential dispersion model function

$$f(y; \mu, \sigma^2) = \exp\left[\frac{1}{\sigma^2}\left\{y\theta(\mu) - K(\theta(\mu))\right\}\right] a(\sigma^2, y),$$

$y \in \mathcal{Y}$, $\mu \in \mathcal{M}$, $\sigma^2 \in \Phi \subseteq (0, \infty)$, and variance function $V(\mu)$. Use the identity $\mu = \mu(\theta(\mu))$ to show that

$$\frac{d\theta}{d\mu} = \frac{1}{V(\mu)}.$$

Verify this identity for the normal, Poisson, $\text{Bin}(1, \mu)$, gamma and inverse Gaussian distributions.

8. Consider a generalised linear model for independent random variables $Y_1, \ldots, Y_n$, with $Y_i \sim ED(\mu_i, \sigma_i^2 V(\mu_i))$, for $i = 1, \ldots, n$ and where $g(\mu_i) = x_i^T \beta$ and $\sigma_i^2 = \sigma^2 a_i$.

   (a) Use the chain rule to show that the likelihood equations for β may be written as
   $$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ir}}{\sigma_i^2 V(\mu_i) g'(\mu_i)} = 0, \quad r = 1, \ldots, p.$$

   (b) Show that the $p \times p$ block of the Fisher information matrix corresponding to β (ignoring the part that depends on σ^2) can be expressed as $i(\beta) = X^T W X$, where X has ith row x_i^T for $i = 1, \ldots, n$ and W is a matrix which you should specify.

   [*Hint: Use the definition of the Fisher information in terms of products of first derivatives of the likelihood function.*]

   (c) How do the expressions in (a) and (b) simplify when $g(\mu_i)$ is the canonical link function?

9. Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim N(\mu_i, \sigma^2)$ and $\mu_i = x_i^T \beta$, for $i = 1, \ldots, n$. Show that the deviance is equal to the residual sum of squares.

10. Return to the `AlloyData` example from Practical Sheet 6. In the output from `summary(BinMod1)`, what is the approximation used to compute the standard errors of the parameter estimates? How are the z-values and the null and residual deviances calculated? Check your answers by doing the calculations in `R` yourself.