5. **Schistosoma japonicum** is a parasitic worm and humans can be infected when using fresh water infested with the infective larval stage of the parasite. Infection is considered endemic in a village if over half of the population are infected. θ is the proportion of the village infected. The investigator randomly samples 11 villagers and measures that 7 of them are infected by the parasite.

(a) What is the likelihood of θ? (Assume that the data has a binomial distribution)

[2 marks]

(b) Assume that the prior for θ is uniform on $[0,1]$. What is the posterior? Which Beta distribution is it (see table 1)?

[3 marks]

(c) What is the maximum likelihood estimate $\hat{\theta}$?

[3 marks]

(d) Show that the standard error of the proportion can be approximated by:

$$\sqrt{\frac{\hat{\theta}\,(1-\hat{\theta})}{n}}$$

[5 marks]

(e) What is the 95% confidence interval of $\hat{\theta}$?

[2 marks]

(f) What is the posterior probability that the infection is endemic?

[5 marks]

(g) After observing the villagers' contact with water, the investigator is 80% sure that the infection is endemic in the village. Explain what new prior you would use to model the investigator's belief.

[5 marks]

(25 marks)

200 ... DW 22/03/16

April 12th ...

6. **Logistic Regression.** Table 2 presents the Beetle mortality data: the variable x_i represents the dose level of toxic substance given to a group i of n_i beetles where k_i beetles died consequently. The likelihood of the proportion θ_i of dead beetles in the group i is modelled using the Binomial distribution.

(a) What is the likelihood of $(\theta_1,\cdots,\theta_8)$?

[3 marks]

(b) What are the maximum likelihood estimates $\hat{\theta}_i, \forall i=1,\cdots,8$?

[3 marks]

(c) A logit model is fitted to the data using the dose level x as the explanatory variable:

$$g(\theta_i)=\log\left(\frac{\theta_i}{1-\theta_i}\right)=\beta_1+\beta_2\,x_i$$

Show that

$$\theta_i=\frac{\exp\left(\beta_1+\beta_2\,x_i\right)}{1+\exp\left(\beta_1+\beta_2\,x_i\right)}$$

[3 marks]

(d) Write the log-likelihood function with respect to the variables β_1 and β_2.

[4 marks]

(e) Give the Newton-Raphson algorithm for estimating β_1 and β_2.

[4 marks]

(f) The R output of this logistic regression is given in table 3. Identify the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ and compute the estimates $\hat{\theta}_1$ and $\hat{\theta}_8$ with this logit model.

[4 marks]

(g) Using the deviance, discuss the goodness of fit of this logistic model.

[4 marks]

(25 marks)

The Beta probability function is defined as:

$$p(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad (1)$$

with $\theta\in[0,1]$, $\alpha>0$ and $\beta>0$. The expectation is $\frac{\alpha}{\alpha+\beta}$ and the variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. For strictly positive integers $n\in\mathbb{N}^*$, the Gamma function, $\Gamma(\cdot)$, is defined as $\Gamma(n)=(n-1)!$ and, $\forall x\in\mathbb{R}$, $\Gamma(x+1)=x\,\Gamma(x)$.

Table 1: Beta distribution.

group	x_i	n_i	k_i
$i=1$	1.6907	59	6
$i=2$	1.7242	60	13
$i=3$	1.7552	62	18
$i=4$	1.7842	56	28
$i=5$	1.8113	63	52
$i=6$	1.8369	59	53
$i=7$	1.8610	62	61
$i=8$	1.8839	60	60

Table 2: Beetle mortality data.


```
Call:
glm(formula = beetle.mat ~ x, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
x             34.270      2.912   11.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

Table 3: R output (logit).

22/03/16 ALSM 2

EXAM PAPER 2010

Q6 A For each group i, the likelihood is:

$$lik(\theta_i) = \binom{n_i}{k_i}\theta_i^{k_i}(1-\theta_i)^{n_i-k_i}$$ (calculate 8 times)

Assuming independence:

$$lik(\theta_1, \dots, \theta_8) = \prod_{i=1}^{8}\left[\binom{n_i}{k_i}\theta_i^{k_i}(1-\theta_i)^{n_i-k_i}\right]$$

B Maximum likelihood estimates of $\forall i$ are given as $\hat{\theta}_i = \frac{k_i}{n_i}$ (max likelihood of saturated model)

C. $g(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_1 + \beta_2 x_i$

$$\theta_i / 1-\theta_i = \exp(\beta_1 + \beta_2 x_i)$$

$$\theta_i = (1-\theta_i)\exp(\beta_1+\beta_2 x_i)$$

$$\theta_i = \exp(\beta_1+\beta_2 x_i) - \theta_i \exp(\beta_1+\beta_2 x_i)$$

$$\theta_i + \theta_i(\exp(\beta_1+\beta_2 x_i)) = \exp(\beta_1+\beta_2 x_i)$$

$$\theta_i(1+\exp(\beta_1+\beta_2 x_i)) = \exp(\beta_1+\beta_2 x_i)$$

$$\theta_i = \frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}$$

D. log likelihood

$$loglik(\theta_1, \dots, \theta_8) = \Sigma_{i=1}^{8}\left[\log\binom{n_i}{k_i} + k_i \log\left(\frac{\theta_i}{1-\theta_i}\right) + n_i \log(1-\theta_i)\right]$$ Saturated model

$$loglik(\beta_1, \beta_2) = \Sigma_{i=1}^{8}\left[\log\binom{n_i}{k_i} + k_i(\beta_1+\beta_2 x_i) + n_i \log(1+\exp(\beta_1+\beta_2 x_i))\right]$$

E. Newton Raphson Method

$$\begin{pmatrix}\beta_1^{m+1} \\ \beta_2^{m+1}\end{pmatrix} = \begin{pmatrix}\beta_1^{m} \\ \beta_2^{m}\end{pmatrix} - \underbrace{H^{-1}_{\beta_1^m \beta_2^m}}_{\text{Hessian of loglik}} \nabla loglik|_{\beta_1^m \beta_2^m}$$

$$\nabla loglik \left| \begin{array}{l} \frac{d\, loglik}{d\beta_1} = \Sigma_{i=1}^{8}\left\{k_i - n_i\left(\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right)\right\} \\ \frac{d\, loglik}{d\beta_2} = \Sigma_{i=1}^{8}\left\{k_i x_i - n_i x_i\left(\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right)\right\} \end{array}\right.$$

Second Derivative:

$$\frac{d^2}{\beta_1^2} = -\Sigma_{i=1}^{8} n_i \theta_i(1-\theta_i) \qquad \theta_i = \frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}$$

$$\frac{d^2}{d\beta_2^2} = -\sum_{i=1}^{8} n_i x_i^2 \theta_i(1-\theta_i)$$

$$\frac{d^2}{d\beta_1 \beta_2} = -\sum_{i=1}^{8} n_i x_i \theta_i(1-\theta_i)$$

Scoring Method Solution

- Expectation of Hessian matrix wrt θ
- y_i's don't appear in second derivative
- Expectation of a constant is a constant
- Expectation of Hessian is the hessian ⇒ same outcome
- Will be different using different link functions etc

F. $\beta_1 = -60.71 \quad \beta_2 = 34.27$

$$\theta_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

$$\hat{\theta}_1 = \frac{\exp(-60.71 + 34.27(1.6097))}{1 + \exp(-60.71 + 34.27(1.6097))} = 0.0584$$

$$\hat{\theta}_5 = \frac{\exp[-60.71 + 34.27(1.8113)]}{1 + \exp[-60.71 + 34.27(1.8113)]} = 0.7932$$

G. Deviance $\sim \chi^2_{(n-m, N=0)} \sim \chi^2_{8-2}$

$D_{calc} = 11.232 \quad D.F = 6 \quad$ Compare to $Dev_{critical}$

Rule of thumb, if less than $2^*(m-n)$ then good model

Value from table is 12.5, to the left ⇒ good model.

24/03/16 ALSM 2

EXAM PAPER 2010

Q5 A likelihood of binomial: $\binom{11}{7}\theta^7(1-\theta)^4$

B Post $\theta = \begin{cases} \frac{like(\theta)}{\int_0^1 like(\theta)} & \text{for } \theta \in [0,1] \\ 0 & \text{Otherwise} \end{cases}$ normalising the likelihood

Post $\theta = \theta^7(1-\theta)^4 \Rightarrow Beta(8,5)$ $n-k+1$, $k+1$

Posterior is proportional to Beta

C Maximum likelihood estimate of θ is $\frac{7}{11} \approx 0.64$

Uncertainty associated with MLE

D $Beta(n-k+1, k+1) \rightarrow Beta(\alpha=8, \beta=5)$

$Variance_{Beta} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{(k+1)(n-k+1)}{(k+1+n-k+1)^2(k+1+n-k+1+1)}$

look at this as $n \to \infty$ (Approximation of limits) $\frac{1}{n} \to 0$ $\frac{k}{n} \to \theta$

$= \frac{n(\frac{k}{n}+\frac{1}{n})n(1-\frac{k}{n}+\frac{1}{n})}{(n+2)^2(n+3)} \approx \frac{n^2\theta(1-\theta)}{n^3} \approx \frac{\theta(1-\theta)}{n}$ variance

$SE = \sqrt{\frac{\theta(1-\theta)}{n}}$

E. 95% CI for θ?

$\theta = 7/11$ $SE(\theta) = \sqrt{\frac{\frac{7}{11}(\frac{4}{11})}{11}} = 0.145$

$7/11 \pm 2 \times 1.45 = (-2.26, 3.53)$

F. Endemic if proportion is over 0.5

$P(\theta \geq 0.5) = \int_{0.5}^1 post(\theta)\,d\theta$

$= \frac{11!}{7!4!}\int_{0.5}^1 \theta^7(1-\theta)^4\,d\theta$

$(1-\theta)^4 = 1-4\theta+6\theta^2-4\theta^3+\theta^4$

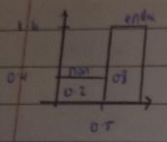
$\theta^7(1-\theta)^4 = \theta^7-4\theta^8+6\theta^9-4\theta^{10}+\theta^{11}$

$= \frac{11!}{7!4!}\int_{0.5}^{1} \theta^7-4\theta^8+6\theta^9-4\theta^{10}+\theta^{11}\,d\theta$

$= \frac{11!}{7!4!}\left[\theta^8/8 - 4\theta^9/9 + \frac{6\theta^{10}}{10} - \frac{4\theta^{11}}{11} + \frac{\theta^{12}}{12}\right]_{0.5}^{1}$

$= 0.8$ (between 0 and 1) $\Rightarrow$ endemic

G Prior Belief



Can use to calculate new prior, part (F) will probably change given this info

$$\mathbb{P}_2 = \frac{\int_{0.5}^{1}\binom{11}{7}\theta^7(1-\theta)^4(1.6)}{\int_0^{0.5}\binom{11}{7}\theta^7(1-\theta)^4(0.4) + \int_{0.5}^{1}\binom{11}{7}\theta^7(1-\theta)^4(1.6)}$$

$$\frac{\binom{11}{7}(1.6)\left[\theta^8/8 - 4\theta^9/9 + \frac{6\theta^{10}}{10} - \frac{4\theta^{11}}{11} + \frac{\theta^{12}}{12}\right]_{0.5}^{1}}{\binom{11}{7}(0.4)\left[\theta^8/8 - 4\theta^9/9 + 6\theta^{10}/10 - 4\theta^{11}/11 + \theta^{12}/12\right]_0^{0.5} + \binom{11}{7}(1.6)\left[\theta^8/8 - 4\theta^9/9 + 6\theta^{10}/10 - 4\theta^{11}/11 + \theta^{12}/12\right]_{0.5}^{1}}$$

$$= \frac{1651/15360}{1.458\times10^{-5} + 1651/15360} = 0.99981$$

Which is $> 0.95 \Rightarrow$ it is endemic