Statistical modelling for biologists and biotechnologists, ST2304

May, 2015

Solution

Problem 1

a) If individuals of the species are trapped at a constant rate λ per unit time, different individuals are trapped independently of each other such that the number of individuals trapped in disjoint intervals are independent, and only one individual can be trapped at a given point in time, then this represents a Poisson process and the total number of individuals X_i trapped during a time interval of length t (24 hours) is Poisson distributed with parameter $\lambda t = 0.002$. To compute this in R using the built in cumulative distribution function `ppois`, we first note that this computes probabilities of the form $P(X > a)$ if the optional argument `lower.tail=FALSE` is given. Noting also that $P(X_i \geq 1) = P(X_i > 0)$, the probability can be computed in R with the expression

```
p.present <- ppois(0,0.002,lower.tail=FALSE)
```

b) The number of traps that we must examine before we find one with the rare species present is geometrically distributed. Most textbooks use the parameterization

$$P(Y = y) = p(1-p)^{y-1} \tag{1}$$

where $y = 1, 2, \ldots$ includes the first success whereas in R, Y' is defined such that it only includes the failures up to the first success and

$$P(Y' = y) = p(1-p)^{y} \tag{2}$$

for $y = 0, 1, \ldots$ (see the help page). If we interpret Y as defined by (1) (both interpretation give a full score here and in point c) then $P(18.2 < Y < 30.5) = P(18.2 < Y' + 1 < 30.5) = P(17.2 < Y' < 29.5)$ can be computed as

```
pgeom(29.5, p.present) - pgeom(17.2, p.present)
```

or, more consisely,

```
diff(pgeom(c(29.5,17.2),p.present))
```

Note also that it makes no difference if we interpret the question such that the endpoints are included in the event since Y can only take integer values.

c) A 95%-probability intervals for Y and Y' can be computed as follows

```
qgeom(c(.025,.975), p.present)
qgeom(c(.025,.975), p.present) + 1
```

Problem 2

a) Lettting y_i and t_i represent the number of fledglings produced by pair i, $i = 1, 2, \dots, n$, the model assumes that each

$$Y_i \sim \text{pois}(\lambda_i) \tag{3}$$

where

$$\ln \lambda_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2. \tag{4}$$

In terms of the expected number of fledglings,

$$E(Y_i) = \lambda_i = e^{\beta_0 + \beta_1 t_i + \beta_2 t_i^2}. \tag{5}$$

b) Rewriting the exponent in the reparameterized model as a second order polynomial in t we see that

$$EY = y_0 e^{-\frac{1}{2}(\frac{t-t_0}{\omega})^2} = e^{\ln y_0 - \frac{t_0^2}{2\omega^2} + \frac{t_0}{\omega^2}t - \frac{1}{2\omega^2}t^2} \tag{6}$$

which equals (4) for all t if

$$\beta_0 = \ln y_0 - \frac{t_0^2}{2\omega^2}, \quad \beta_1 = \frac{t_0}{\omega^2}, \quad \beta_2 = -\frac{1}{2\omega^2}. \tag{7}$$

The width of the Gauss-curve ω as function of β_2 is thus

$$\omega = \sqrt{-\frac{1}{2\beta_2}} \tag{8}$$

c) Let

$$\omega = f(\beta_2) = \frac{1}{\sqrt{2}}(-\beta_2)^{-1/2}. \tag{9}$$

Functional invariance of MLEs implies that the MLE of ω is

$$\hat{\omega} = f(\hat{\beta}_2) = \sqrt{-\frac{1}{2\hat{\beta}_2}} = \sqrt{-\frac{1}{2(-0.00108)}} = 21.49 \tag{10}$$

days. To use the delta methods, we need

$$\frac{df}{d\beta_2} = -\frac{1}{2\sqrt{2}}(-\beta_2)^{-3/2}(-1) = \frac{1}{2\sqrt{2}}0.00108^{-3/2} = 9961 \tag{11}$$

at $\beta_2 = \hat{\beta}_2 = -0.00108$. The variance and standard error of $\hat{\omega}$ is then

$$\operatorname{Var}\hat{\omega} = \left(\frac{df}{d\beta_2}\right)^2 \operatorname{Var}(\beta_2) = (9961)^2 0.0004782^2 = 22.69, \tag{12}$$

and

$$\operatorname{SE}\hat{\omega} = 4.76 \tag{13}$$

days.

d) Under the null hypothesis of no overdispersion $H_0 : \varphi \leq 1$, the deviance is chi-square with $n - p = 92 - 3 = 89$ degrees of freedom. We reject the null hypothesis in favour of the hypothesis of overdispersion ($H_1 : \varphi > 1$) for large values of $D > \chi^2_{\alpha,89} \approx \chi^2_{0.05,90} = 113.14$. Since the observed value of $D = 353.81$ is greater than the critical value we reject H_0 in favour of H_1.

Given that there is overdispersion in the data, the standard error in point c based on the assumption of no overdispersion underestimates the true standard error. Given the estimate of the dispersionparameter $\hat{\varphi} = 2.8$, adjusted estimates of the standard errors of $\hat{\beta}_2$ and $\hat{\omega}$ become $\operatorname{SE}(\hat{\beta}_2) = \sqrt{2.8} \cdot 0.000482 = 0.00800$ and $\operatorname{SE}(\hat{\omega}) = \sqrt{2.8} \cdot 22.7 = 38.0$.

Mechanisms which may lead to overdispersion are missing covariates (for example, variables characterizing the environment surrounding each nestbox, variables characterizing the phenotypes of the parents), positive covariances between the survival of individual offspring, and an incorrect choice of link function.

e) From the output form the call to `optim` and the computed inverse hessian matrix at the maximum we find $\hat{p}_0 = 0.271$, $\operatorname{SE}\hat{p}_0 = 0.046$, $\hat{\lambda} = 6.74$, and $\operatorname{SE}\hat{\lambda} = 0.32$.

f) Without zero-inflation, each Y_i is Poisson distributed with parameter λ and the likelihood becomes

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\sum y_i}}{\prod y_i!} \tag{14}$$

and the log likelihood

$$\ln L(\lambda) = -n\lambda + \ln\lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!). \tag{15}$$

This is maximised for $\hat{\lambda} = \bar{y} = \frac{1}{n}\sum y_i$ at which

$$\ln L(\hat{\lambda}) = -92 \cdot 4.91 + \ln 4.91 \cdot 452 - 570.78 = -303.25 \tag{16}$$

To test the null hypothesis of no zero-inflation (H_0) against the model including zero-inflation (H_1) we note that these models are nested since H_0 can be seen as a special case ($p_0 = 0$) of H_1. Using $2(\ln L_1 - \ln L_0)$ which is approximately chi-square with $p_1 - p_0 = 2 - 1 = 1$ degrees of freedom under H_0, the observed value of the test statistic becomes $2(-213.65 - (-303.25)) = 180$ which is higher than the critical value $\chi^2_{0.05,1} = 3.84$. Thus we reject H_0.[1]

g) Several solutions are possible. The $I_0(y)$ part of the expression for $P(Y = y)$ which needs to be computed for all observation, can be represented in R by an the vectorized conditional selection statement `ifelse(y==0,1,0)`. This becomes a vector taking values of 1 for each element of `y` for all other elements. Using this the probabilities of each observation can be computed in a single expression. The total log likelihood is then found by taking logs and summing over all observations as follows.

```
lnL <- function(par,y) {
  p0 <- par[1]
  lambda <- par[2]
  -sum(log(p0*ifelse(y==0,1,0) + (1-p0)*dpois(y,lambda)))
}
```

An alternative solution is to first compute all probabilities based on the second term only and then use a logical vector to add the zero-inflation probabilities to all observations for which y is zero.

```
lnL <- function(par,y) {
  p0 <- par[1]
  lambda <- par[2]
  p.obs <- (1-p0)*dpois(y,lambda))
  p.obs[y==0] <- p.obs[y==0] + p0
  -sum(log(p.obs))
}
```

[1] In the exam text $\sum_{i=1}^{n} \ln(y_i!)$ was incorrectly given as 452 (instead of the correct value of 570.78 used above) so this point will be marked leniently. If using the incorrect value given, we get $\ln L(\hat{\lambda}) = -184.46$ and $2(\ln L_1 - \ln L_0) = 2(-213.65 - (-184.46)) = -58.38$ which is impossible since H_0 is nested in H_1. Provided that the derivation of the maximum log likelihood under H_0 is correct, a full score is given, regardless of which conclusion for the hypothesis test is reached.