Discrete Multivariate Analysis - 579 - HW #9

Alex Towell (atowell@siue.edu)

Refer to the calf infection study discussed in lecture. Let n_3 denote the number of calves with primary, secondary, and tertiary infections, n_2 the number with only primary and secondary infections, n_1 the number with only a primary infection, and n_0 the number that were never infected. Let θ be the probability of a primary infection. Consider a model where the rate of infection is constant. That is,

$$\Pr(\text{tertiary}|\text{secondary}, \text{primary}) = \Pr(\text{secondary}|\text{primary}) = \Pr(\text{primary}).$$

Problem 1

Derive the maximum likelihood estimator of θ under the constant infection rate model.

We assume the data model is given by

$$(n_0, n_1, n_2, n_3) \sim \text{MULT}(n, \pi_0, \pi_1, \pi_2, \pi_3)$$

where n_j denotes the number of j infections observed and π_j denotes the probability of j infections.

First, we assume $\Pr(\text{primary}) = \theta$. Then, given that in order for the secondary infection to occur, the primary infection must occur and in order for the tertiary infection to occur, the secondary infection must occur, consider the following.

Under the constant infection rate model, the probability of 0 infections is the probability that the primary infection does not occur, $\pi_{00}(\theta) = 1 - \theta$. The probability of 1 infection is the probability the primary infection occurs and the secondary infection does not occur, $\pi_{10}(\theta) = \theta(1-\theta)$. The probability of 2 infections is the probability the primary and secondary infections occur and the tertiary infection does not occur, $\pi_{20}(\theta) = \theta^2(1-\theta)$. Finally, the probability of 3 infections is the probability the primary, secondary, and tertiary infections occur, $\pi_{30}(\theta) = \theta^3$.

Thus,

$$\{n_j\} \sim \text{MULT}(n, \{\pi_{j0}(\theta)\}).$$

Under the null model, the likelihood function is given by

$$\text{L}(\theta) \propto \prod_{j=0}^{3} \pi_{j0}(\theta)^{n_j}$$

and the kernel of the log-likelihood function is given by

$$\begin{aligned}
\ell(\theta) &= \sum_{j=0}^{3} n_j \log \pi_{j0}(\theta) \\
&= n_0 \log \pi_{00}(\theta) + n_1 \log \pi_{10}(\theta) + n_2 \log \pi_{20}(\theta) + n_3 \log \pi_{30}(\theta) \\
&= n_0 \log(1-\theta) + n_1 \log \theta(1-\theta) + n_2 \log \theta^2(1-\theta) + n_3 \log \theta^3 \\
&= n_0 \log(1-\theta) + n_1 \log \theta + n_1 \log(1-\theta) + 2n_2 \log \theta + n_2 \log(1-\theta) + 3n_3 \log \theta \\
&= (n_1 + 2n_2 + 3n_3) \log \theta + (n_0 + n_1 + n_2) \log(1-\theta).
\end{aligned}$$

The derivative of the log-likelihood is thus given by

$$\frac{d\,\ell}{d\theta} = \frac{n_1 + 2n_2 + 3n_3}{\theta} - \frac{n_0 + n_1 + n_2}{1-\theta}.$$

The ML estimate of θ is given by

$$\left.\frac{d\,\ell}{d\theta}\right|_{\hat{\theta}} = 0.$$

Solving for $\hat{\theta}$, we see that

$$\frac{n_1 + 2n_2 + 3n_3}{\hat{\theta}} = \frac{n_0 + n_1 + n_2}{1-\hat{\theta}}.$$

Next, we take the reciprocal of both sides,

$$\frac{\hat{\theta}}{n_1 + 2n_2 + 3n_3} = \frac{1-\hat{\theta}}{n_0 + n_1 + n_2}.$$

Expanding the RHS, we get

$$\frac{\hat{\theta}}{n_1 + 2n_2 + 3n_3} = \frac{1}{n_0 + n_1 + n_2} - \frac{\hat{\theta}}{n_0 + n_1 + n_2}.$$

Adding $\frac{\hat{\theta}}{n_0+n_1+n_2}$ to both sides,

$$\hat{\theta}\left(\frac{1}{n_1 + 2n_2 + 3n_3} + \frac{1}{n_0 + n_1 + n_2}\right) = \frac{1}{n_0 + n_1 + n_2}.$$

Dividing both sides by the part in parantheses, we get

$$\hat{\theta} = \frac{1}{(n_0 + n_1 + n_2)\left(\frac{1}{n_1+2n_2+3n_3} + \frac{1}{n_0+n_1+n_2}\right)},$$

which simplifies to

$$\hat{\theta} = \frac{n_1 + 2n_2 + 3n_3}{n_0 + 2n_1 + 3n_2 + 3n_3}.$$

Problem 2

We observe data $n_0 = 63$, $n_1 = 63$, $n_2 = 25$, $n_3 = 5$. Compute the mle $\hat{\theta}$ and the estimated probabilities $\vec{\pi}_0(\hat{\theta})$.

Using the MLE equation, we compute the result using R:

```
n0 <- 63
n1 <- 63
n2 <- 25
n3 <- 5
mle <- (n1 + 2*n2 + 3*n3)/(n0 + 2*n1 + 3*n2 + 3*n3)
mle
```

```
## [1] 0.4587814
```

Thus, we see that

$$\hat{\theta} = 0.4587814.$$

Plugging in this result,

$$\vec{\pi}_0(\hat{\theta}) = (1-\hat{\theta}, \hat{\theta}(1-\hat{\theta}), \hat{\theta}^2(1-\hat{\theta}), \hat{\theta}^3)$$

which we compute with R:

```r
pihats <- c(1-mle,mle*(1-mle),mle^2*(1-mle),mle^3)
round(pihats,digits=3)
```

```
## [1] 0.541 0.248 0.114 0.097
```

Thus, we see that $\vec{\pi}_0(\hat{\theta}) = (0.541, 0.248, 0.114, 0.097)'$.

Problem 3

Part (a)

Compute the statistic G^2 and the p-value for testing the constant infection rate model.

We use the following R code to compute the observed statistic G_0^2.

```r
# below is code for testing a specified multinomial
# data is entered as a vector of counts
obs = c(n0,n1,n2,n3)
n = sum(obs)

# the null model is entered as a vector of hypothesized probabilities
# expected counts under the null model are computed as n*prob
m = n*pihats

# log-likelihood ratio test
G2 = 2*sum(obs*log(obs/m))
G2
```

```
## [1] 30.43096
```

We see that $G_0^2 = 30.43096$.

The reference distribution is the chi-squared distribution with df $= c - 1 - 1 = 2$ degrees of freedom. The following R code computes the p-value:

```r
pchisq(G2,2,lower.tail=FALSE)
```

```
## [1] 2.466041e-07
```

We see that p-value < 0.001.

Part (b)

Provide an interpretation of your result, stated in the context of the problem.

We obtained a p-value less than 0.001, which we consider to be very strong evidence against the constant infection rate model.

Problem 4

Part (a)

Compute $\hat{\pi}$, the estimated probability vector under the full model.

Recall that the data model is given by

$$(n_0, n_1, n_2, n_3) \sim \text{MULT}(n, \pi_0, \pi_1, \pi_2, \pi_3).$$

Under the full model, each π_j may be independently specified.

We simply let the data speak for itself and estimate π_j as

$$\hat{\pi}_j^F = \frac{n_j}{n}.$$

We use the following R code to compute the probability vector:

```
pihats_full = obs/n
round(pihats_full,digits=3)
```

```
## [1] 0.404 0.404 0.160 0.032
```

Part (b)

Compute the estimated conditional probabilities of further infection under the full model.

The probability of a primary infection is

$$\Pr(\text{primary}) = \pi_1 + \pi_2 + \pi_3,$$

the probability of a secondary infection given a primary infection is

$$\Pr(\text{secondary} \,|\, \text{primary}) = \frac{\pi_2 + \pi_3}{\pi_1 + \pi_2 + \pi_3},$$

and the probability of a tertiary infection given a secondary infection is

$$\Pr(\text{tertiary} \,|\, \text{secondary}) = \frac{\pi_3}{\pi_2 + \pi_3}.$$

We estimate these probabilities by plugging in the MLEs:

```
primary = sum(pihats_full[2:4])
secondary = sum(pihats_full[3:4]) / primary
tertiary = pihats_full[4] / secondary

round(c(primary,secondary,tertiary),digits=3)
```

```
## [1] 0.596 0.323 0.099
```

We see that the conditional probability vector is given by $(0.596, 0.323, 0.099)'$.

Part (c)

Provide an interpretation, stated in the context of the problem.

The conditional probabilities for the infection rates are significantly *decreasing*, i.e., not constant. The constant infection rate model was not able to capture this relationship in the data.