ML estimation - A fictitious paper -

FPA2020 Lecture 6

Taro Kaiyo, Hanako Kaiyo and Toshihide Kitakado* (TUMSAT)

June 3, 2020

Contents

1 INTRODUCTIOIN

The aim of this paper is to analyze data of school size for xxx species. In the previous studies, the mean school size in this population was estimated as 6. However, due to a recent change in the habitat condition, it has been suggested that the mean school size might be decreased to some extent. Therefore, we conducted a field experiment to observe the school size for this species.

During the experiment, it was concerned that the observed school size may be subject to "size bias", which means that larger the school size is, higher the detection probability might be. This means, a simple average of school size data tends to be positively biased.

In this paper, we estimate the means schoool size using the likelihood method. For this purpose, we constructed two statistical models to account for the size bias in the observation process. We then compared these models with a model with no size bias. We also investigate if the previous knowledge on the mean school size is still correct or not.

2 MATERIALS and METHODS

2.1 Observed data

We observed the following school sizes from a total of 58 detected schools through the experiment:

```
 [1]  6  4  3  6  6  7  4  5  9  6  4  3  7  7  3  6  5  9  9  8  4  8  5 10  3
[26]  5  5  5  6  7  5  8  7  5  3  5  3  5  4 11  5  6  3  7  4  5  4  7  3  4
[51]  3  4  3  5  7  6  6  4
```

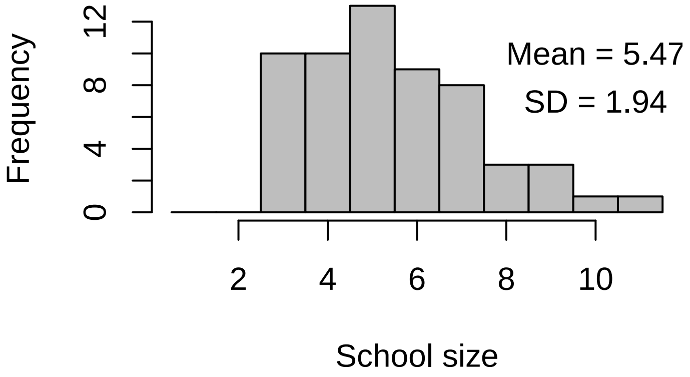

Figure 1: Histogram of observed school sizes

2.2 Statistical models

2.2.1 Conventional shifted poisson model (Model 1)

For a model without consideration of any size bias, we employed a shifted poisson distribution as follows:

$$S_1 - 1, S_2 - 1, \dots, S_n - 1 \sim (iid)Po(\lambda).$$

The probability distrbution is shown as

$$f_1(s) = P(S_i = s) = e^{-\lambda}\frac{\lambda^{s-1}}{(s-1)!} \quad (s = 1, 2, \dots).$$

In this model, the expectation of S_i is

$$E[S_i] = \lambda + 1$$

and therefore $\gamma = \lambda + 1$ is the mean school size. So $\lambda = 5$ is the knowledge prior to this analysis.

2.2.2 Size-biased model (Model 2)

Here, we use the following relatively simple size bias model:

$$p(s|\beta) = 1 - e^{-\beta s},$$

where s is the true school size and $\beta(> 0)$ is a parameter.

Under the assumption of size bias, the observed school sizes are those only for schools detected. In this regard, we prepare for a random variable indicating the outcome of detection as

$$I_i = \left\{ \begin{array}{ll} 1 & \text{if the school is detected} \\ s0 & \text{if the school is not detected} \end{array} \right. \quad (i = 1, 2, ..., m)$$

where m is the number of schools which the observer actually encounterd. The probability distribution for I_i is shown below:

$$P(I_i = 1|S = s) = 1 - P(I_i = 0|S = s) = p(s|\beta).$$

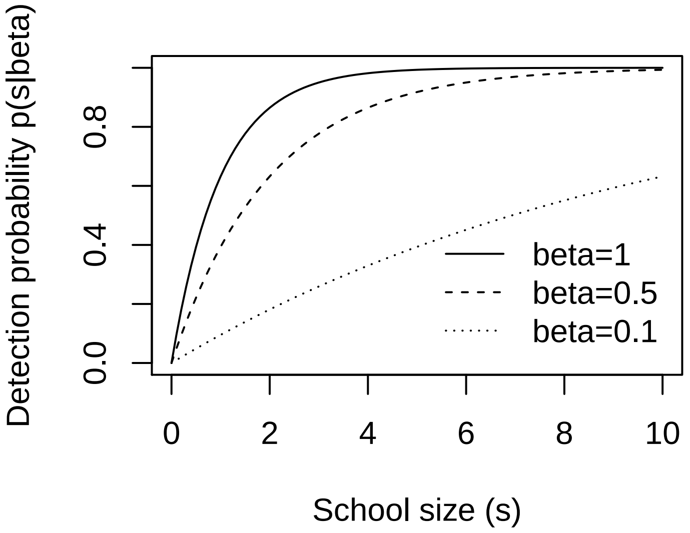

Figure 2: Graphs of $p(s|\beta)$ for β =1, 0.5 and 0.1

Under the setting above, the probability distribution of the observed school size can be expressed by

$$\begin{aligned} P(S_i = s|I_i = 1) &= \frac{P(S_i = s)P(I_i = 1|S_i = s)}{P(I_i = 1)} \\ &= \frac{P(S_i = s)P(I_i = 1|S_i = s)}{\sum_{s=1}^{\infty} P(S_i = s)P(I_i = 1|S_i = s)} \\ &= \frac{e^{-\lambda}\frac{\lambda^{s-1}}{(s-1)!}(1-e^{-\beta s})}{\sum_{s=1}^{\infty} e^{-\lambda}\frac{\lambda^{s-1}}{(s-1)!}(1-e^{-\beta s})} \\ &= \frac{\frac{\lambda^{s-1}}{(s-1)!}(1-e^{-\beta s})}{e^{\lambda} - e^{-\beta}e^{\lambda e^{-\beta}}} \quad (s = 1, 2, ...). \end{aligned}$$

Since $\sum_{i=1}^{m} I_i = n$, and $m-n$ schools among n schools were not detected, we relabel the observed school size as again $S_1, S_2, ...S_n$ as in the no size-biased model and define

$$f_2(s) = \frac{\frac{\lambda^{s-1}}{(s-1)!}(1-e^{-\beta s})}{e^{\lambda} - e^{-\beta}e^{\lambda e^{-\beta}}}$$

2.2.3 Another simple size-biased model (Model 3)

We consider another simple weighted distribution as follows:

$$\begin{aligned} f_3(s) &= \frac{sP(S_i = s)}{\displaystyle\sum_{s=1}^{\infty} sP(S_i = s)} \\ &= \frac{\dfrac{s\lambda^{s-1}}{(s-1)!}}{\displaystyle\sum_{s=1}^{\infty} \dfrac{s\lambda^{s-1}}{(s-1)!}} \\ &= \frac{1}{(1+\lambda)e^{\lambda}} \frac{s\lambda^{s-1}}{(s-1)!} \quad (s = 1, 2, \dots) \end{aligned}$$

This model has only one parameter λ but seems to be robust.

2.3 Statistical inference

2.3.1 Point estimation and the evaluation of standard error

For the estimation of parameter(s), we used the maximum likelihood method. Here, we generally describe the the likelihood function for Model $h(= 1, 2, 3)$ as

$$L_h(\theta) = \prod_{i=1}^{n} f_h(s_i),$$

where θ is a d-dimensional parameter vector ($\theta = (\theta_1, ..., \theta_d)$) and $\theta = \lambda$ for Models 1 and 3, and $\theta = (\lambda, \beta)$ for Model 2.

The point estimate of θ is the maximizer of $\log L(\theta)$ as

$$\hat{\theta} = \arg\ \max_{\theta} \log L(\theta).$$

The standard error of the estimate is assessed via the observed Fisher Information matrix, which is defined like

$$I_{obs}(\hat{\theta}) = -\frac{\partial^2}{\partial\theta\partial\theta^t} \log L(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^t} \log f_h(s_i; \theta),$$

where $\frac{\partial^2}{\partial\theta\partial\theta^t}$ is a matrix operator as

$$\frac{\partial^2}{\partial\theta\partial\theta^t} = \begin{pmatrix} \dfrac{\partial^2}{\partial\theta_1^2} & \cdots & \dfrac{\partial^2}{\partial\theta_1\partial\theta_d} \\ \cdots & & \\ \dfrac{\partial^2}{\partial\theta_d\partial\theta_1} & \cdots & \dfrac{\partial^2}{\partial\theta_d^2} \end{pmatrix}.$$

Once we get $I_{obs}(\hat{\theta})$, we can provide the asymptotic variance as

$$\Sigma = V[\hat{\theta}] \approx I_{obs}^{-1}(\hat{\theta})$$

and the standard error is given by

$$SE[\hat{\theta}_j] = \sqrt{\Sigma_{jj}}.$$

If a paramter transformation $\theta = \exp(\log\theta)$ is used, the standard error is assessed by the delta method as

$$SE[\hat{\theta}_j] = \hat{\theta}\sqrt{\Sigma_{jj}}.$$

because of $d\theta/d\log\theta = \theta$.

2.3.2 Model selection

For comparing the two models, we use Akaike's information criterion, which is defined in general as follows:

$$AIC = -2\log L(\hat{\theta}) + 2d.$$

2.3.3 Statistical test

To test if the information on the current mean school size $\gamma = 6$ $(\lambda = \gamma - 1 = 5)$ is consistent with the previous knowledge, we conduct the likelihood ratio test depending on the selected model as

$$H_0 : \lambda = 5 \quad vs \quad H_1 : \lambda \neq 5$$

with the properties of

$$LRT = -2\log \frac{L_h(\lambda_0)}{L_h(\hat{\lambda})} \to \chi_1^2 \quad \text{under Model } h = 1, 3$$

or

$$LRT_2 = -2\log \frac{L_2(\lambda_0, \beta(\lambda_0))}{L_2(\hat{\lambda}, \hat{\beta})} \to \chi_1^2 \quad \text{under Model 2,}$$

where $\lambda_0 = 5$.

3 RESULTS and DISCUSSION

[From here, I will show my codes for your reference]

3.1 Estimation results

3.1.1 For Model 1

[Actually, for Model 1, we don't need the optimization below, but for the parallel story telling, I used "optim" intentionally.]

```r
NLL.m1 <- function(par){
  lam <- exp(par[1])
  obj <- (-1.0)*sum(dpois(Sobs-1,lam,log=T)) ## Sorry "Sobs" should have been "Sobs-1"
  obj
}
Res.m1 <- optim(1, NLL.m1, method="BFGS", hessian=T)

lam.m1.est <- exp(Res.m1$par[1])
FI.m1 <- Res.m1$hessian
InvFI.m1 <- solve(FI.m1)
lam.m1.se <- lam.m1.est*sqrt(InvFI.m1[1,1])
AIC.m1 <- 2*Res.m1$value + 2*1

data.frame(lam.m1.est, lam.m1.se, AIC.m1)
```

```
  lam.m1.est lam.m1.se  AIC.m1
1   4.465516 0.2774737 239.234
```

You can find that the estimate is same as $\bar{S} - 1 = 4.465$. It is also shown that the Wald confidence interval of λ includes 5, so in the sense of Wald test, the null hypothesis is accepted.

```
zz <- qnorm(0.975); zz
```

```
[1] 1.959964
```

```
data.frame(LB=lam.m1.est-zz*lam.m1.se, UB=lam.m1.est+zz*lam.m1.se)
```

```
        LB       UB
1 3.921677 5.009354
```

3.1.2 For Model 2

```
NLL.m2 <- function(par){
  lam <- exp(par[1])
  beta <- exp(par[2])
  numerator <- lam^(Sobs-1)*(1-exp(-beta*Sobs))/gamma(Sobs)
  denominator <- exp(lam)-exp(-beta+lam*exp(-beta))
  obj <- (-1.0)*sum(log(numerator)-log(denominator))
  obj
}

init <- log(c(lam.true,beta.true))
Res.m2 <- optim(init, NLL.m2, method="BFGS", hessian=T, control=list(reltol=10^(-10)))
lam.m2.est <- exp(Res.m2$par[1])
beta.m2.est <- exp(Res.m2$par[2])

FI.m2 <- Res.m2$hessian
InvFI.m2 <- solve(FI.m2)
lam.m2.se <- lam.m2.est*sqrt(InvFI.m2[1,1])
beta.m2.se <- beta.m2.est*sqrt(InvFI.m2[2,2])
AIC.m2 <- 2*Res.m2$value + 2*2

data.frame(lam.m2.est, lam.m2.se, beta.m2.est, beta.m2.se, AIC.m2)
```

```
  lam.m2.est lam.m2.se beta.m2.est beta.m2.se   AIC.m2
1   3.698765  1.534374  0.01069984  0.8364867 239.3646
```

As shown above, the level of estimation uncertainty is huge for both the parameters. Also, a slightly larger AIC is obtained. This can be confirmed with contour plots of loglikelihood with respect to the parameters. The model is well-defined, but it seems that the data do not have enough information to estimate β well.

```
Len <- 100
LL <- array(0, c(Len,Len))
loglamvec <- seq(1,2,length.out=Len)
logbetavec <- seq(-10,0,length.out=Len)
for(i in 1:Len){
  for(j in 1:Len){
    LL[i,j] <- NLL.m2(c(loglamvec[i], logbetavec[j]))
  }
}
par(mfrow=c(1,2))
contour(loglamvec,logbetavec,LL,nlevels=30, xlab="log(lambda)", ylab="log(beta)")
contour(exp(loglamvec),exp(logbetavec),LL,nlevels=100, xlab="lambda", ylab="beta")
```

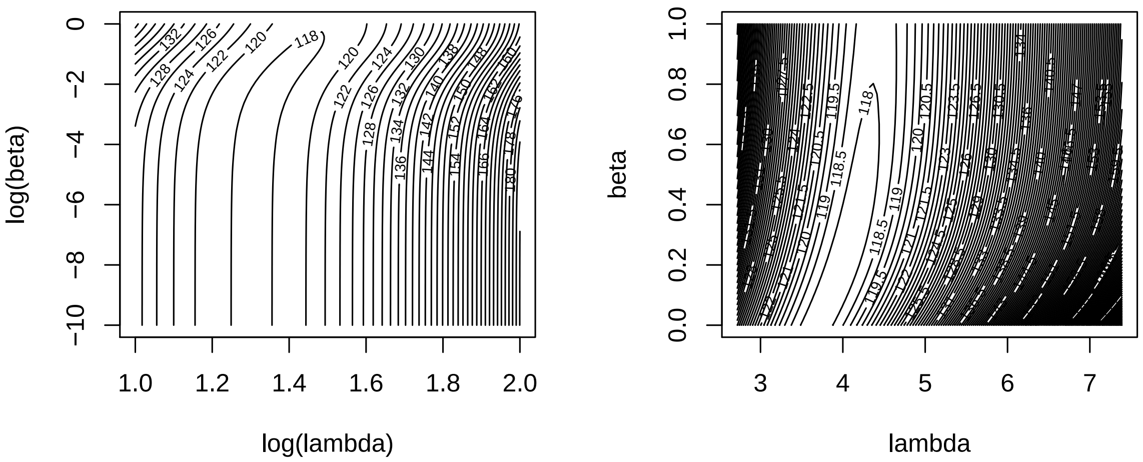

Figure 3: Likelihood contour for Model 2

3.1.3 For Model 3

```
NLL.m3 <- function(par){
  lam <- exp(par[1])
  numerator <- lam^(Sobs-1)*Sobs/gamma(Sobs)
  denominator <- (1+lam)*exp(lam)
  obj <- (-1.0)*sum(log(numerator)-log(denominator))
  obj
}

init <- log(c(lam.true))
Res.m3 <- optim(init, NLL.m3, method="BFGS", hessian=T, control=list(reltol=10^(-10)))
lam.m3.est <- exp(Res.m3$par)

FI.m3 <- Res.m3$hessian
InvFI.m3 <- solve(FI.m3)
lam.m3.se <- lam.m3.est*sqrt(InvFI.m3)
AIC.m3 <- 2*Res.m3$value + 2*1

data.frame(lam.m3.est, lam.m3.se, AIC.m3)
```

```
  lam.m3.est lam.m3.se   AIC.m3
1   3.679228 0.2463012 237.3645
```

We produced a similar estimate of λ with Model 2, but we obtain it with a better precision and a smaller AIC value. Also, it is obvious that the Wald confidence interval does not include 5, so in the sense of Wald test, the null hypothesis is rejected.

3.2 Model selection

Based on the values of AIC, Model 3 was selected. In this model, the estimate of λ is 3.68 with SE=0.25, and the null hypothesis $H_0 : \lambda = 5$ is rejected in the sense of Wald test.

3.3 Likelihood ratio test

We conducted the likelihood ratio test under Models 2 and 3.

3.3.1 Under Model 2

```
NLL.prof.m2 <- function(par, lambda){
  lam <- lambda
  beta <- exp(par[1])
  numerator <- lam^(Sobs-1)*(1-exp(-beta*Sobs))/gamma(Sobs)
  denominator <- exp(lam)-exp(-beta+lam*exp(-beta))
  obj <- (-1.0)*sum(log(numerator)-log(denominator))
  obj
}

Res.H0.m2 <- optim(log(beta.true), NLL.prof.m2, lambda=5, method="BFGS", hessian=T)
LRT <- (-2)*(Res.m2$value - Res.H0.m2$value)
LRT
```

```
[1] 5.040245
```

In Model 2, the value LRT was 5.04, which is greater than the critical value of 3.84, and therefore again H_0 is rejected. This result is contradictory to the result of Wald test in this model, but it seems that the LRT test might be trustable. However, we need to leave the decision for the LRT result under Model 3.

3.3.2 Under Models 1 and 3

```
LRT_1 <- -2*(-NLL.m1(log(5))+NLL.m1(Res.m1$par)); LRT_1
```

```
[1] 3.438618
```

```
LRT_3 <- -2*(-NLL.m3(log(5))+NLL.m3(Res.m3$par)); LRT_3
```

```
[1] 23.16149
```

As easily expecetd, the LRT test under Model 1 wrongly supports the null hypothesis, but the best model (Model 3) rejected it.

4 CONCLUSION

- Actually, I set the true model as Model 2 with parameters $\lambda = 4$ (mean school size $= \gamma = 5$) and $\beta = 0.2$. So, we can find that the estimate under Model 1, $\hat{\lambda} = 4.47$ ($95\%CI = 3.92 - 5.02$) is overestimated as we thought. Also, the null hypothesis, $\lambda = 5$, was wrongly accepted.
- In the true Model 2, it was difficult to estimate the parameters precisely even when assuming the true model.
- On the contrary, Model 3 is simple and not the true model, but practically performed well for the estimation and testing, and the AIC supported this model!

Acknowledgement

Thanks to M1 students in my lab for their swift check!