

Homework # 2, October 8, 2009

Due: In class, 22 Oct 2009.

Reminders:

1. Choose **4 of the 5** problems. You are not expected to do all five.
2. Please see the **Homework guidelines** page of the class web site for homework policies.
3. In particular, you are encouraged to work together. However, please write up your own answers.
4. This is not a programming class. I will not provide functions (except for some of problem 2), but I will certainly help you get your functions working.
5. None this HW's problems need the "data problem" style answer. The data problems this time are designed to make specific points about methods.

1. Last HW, we looked at data from cottontail rabbits. You should have found that the M0, Mt, and Mb models had similar lnL values. Model averaging can be really useful when the choice of model isn't obvious.
 - (a) Calculate the model averaged estimate of the population size and its variance using AIC (or AICc if you prefer) weights.
 - (b) An alternative approach to model averaging is based on BIC. Again, you calculate ΔBIC for each model and $w_i = \exp(-\Delta\text{BIC}/2)$. The weights for each model are $w_i/\sum w_i$. These weights have some theoretical support as approximate posterior model probabilities. Repeat the model averaging using BIC weights.

2. The topic of model averaging led to some discussion in class and some scepticism as to its value. The following problem is meant to explore the benefit (or lack thereof) of model averaging.

The biological question, 'is model averaging a good idea?' has many possible aspects. One consideration is statistical: does model averaging lead to estimates (of N) that are closer to the true value? When the estimates are obtained by numerical maximization, the easiest way to answer the statistical question is to use simulation. This requires specifying a model for the population. Let us consider two cases. The first is that the true model is Mt with parameters: $N = 100$, $p_1=0.3$, $p_2=0.35$, $p_3=0.4$. As you should realize, we're considering a study with 3 capture occasions.

Please:

simulate a set of data from this model

fit each of the three 'easy' models (M0, Mt, and Mb), compute AIC and BIC

Consider four estimators of N :

\hat{N} from the best fitting model using AIC

\hat{N} from the best fitting model using BIC

The model averaged estimate using AIC (or AICc) weights

The model averaged estimate using BIC weights

Record each estimate of N

repeat for 200 randomly generated data sets

Then compute the bias (average $\hat{N} - 100$) and mean-squared error (average $(\hat{N} - 100)^2$) for each estimator.

Which method is better (at least on average)?

Do this again when the true model is Mt with parameters: $N = 100$, $p_1=0.3$, $p_2=0.5$, $p_3=0.6$.

The file `simmulti.txt` has R functions to:

simulate data from this population and produce a capture history, and

compute sufficient statistics from a capture history.

For last HW, I provided R functions to fit each model from sufficient statistics. I strongly suggest using a loop, e.g. `for ()`, to do this 200 times. (Since these were written at different times, I don't promise that the order of sufficient statistics in the data vector is consistent. Do check!).

If you work together on this problem, please each run a separate set of 200 data sets. That way, if everyone does this problem, we get an answer based on 1400 (more or less) random data sets.

3. In class, I gave you formulae for the Jolly-Seber estimates of survival probability. These are the same as the Cormack-Jolly-Seber estimates of survival probability for marked individuals. Consider an open population with 3 capture occasions. This problem has some algebra but it makes some interesting points about J-S (and C-J-S) estimates.

Remember that the Cormack-Jolly-Seber approach conditions on the occasion of first capture. There are 8 possible capture histories for 3 capture occasions, although one is irrelevant for C-J-S and one is uninformative. Here is my suggested notation:

Capture occasion			obs. #
1	2	3	animals
Y	Y	Y	x_1
Y	Y	N	x_2
Y	N	Y	x_3
Y	N	N	x_4
N	Y	Y	x_5
N	Y	N	x_6
N	N	Y	irrelevant
N	N	N	never marked

- (a) For each of the 6 relevant capture histories, determine the probability of observing that capture history, in terms of the survival probabilities (ϕ_i) and capture probabilities (p_i).
- (b) I gave you equations for all the estimates of J-S parameters. Many were involved \hat{M}_i , but you now know how to estimate \hat{M} . The estimates of \hat{p}_2 and $\hat{\phi}_1$ can be written in terms of some of the x 's from the above set of capture histories. Show that

$$\hat{p}_2 = x_1/(x_1 + x_3), \text{ and}$$

$$\hat{\phi}_1 = \frac{(x_1 + x_2)(x_1 + x_3)}{x_1(x_1 + x_2 + x_3 + x_4)}$$

- (c) I also said that some parameters were not estimable (e.g. ϕ_2 and ϕ_3 for data from 3 capture occasions). Look at the equation for $\hat{\phi}_2$ (I may have omitted the hat in my lecture) and tell me why it is not estimable.
- (d) In the C-J-S approach, you only consider the marked individuals, so \hat{p}_i is estimated by m_i/\hat{M}_i . Is it possible to estimate p_3 ?
- (e) As we'll see in the next part, estimates of ϕ_1 depend on ϕ_2 and p_3 . However, the only place either occur is as the product $\phi_2 p_3$. Is it possible to estimate the product?
- (f) Work out the mle of ϕ_1 , the survival probability between capture occasion 1 and capture occasion 2, by maximizing lnL with respect to ϕ_1 . Along the way, $\hat{\phi}_1$ will depend on other parameters (e.g. p_2 and $\phi_2 p_3$). Substitute the expressions you derived above and simplify. You should end up with the estimator given in class.

4. The data in `madeup.csv` are data that I constructed that may have heterogeneity in capture probabilities among individuals. This population can be considered closed over the 3 capture occasions. For each captured animal, I give you the number of times it was captured (out of 3 possible occasions), its weight, and its age.

- (a) Use model M0 to estimate the population size and the overall capture probability. Use profile likelihood to estimate a confidence interval for N .
- (b) Since you have two covariates, weight and age, for each observed animal, you can use Huggins's method for model Mh (heterogeneity among individuals). Use a standard package for logistic regression to estimate p_i for each individual using both weight and age.

I'm very happy to help here if you're not familiar with logistic regression. Proc logistic in SAS or glm() in R are both very effective. If you know program MARK, you're welcome to use that.

Estimate the population size, accounting for heterogeneity in capture probability. (No variance estimate required, although if you use MARK, please tell what it reports as the variance).

- (c) For these data, would be appropriate to ignore the heterogeneity among animals? Explain why or why not.

Note: I know of at least three ways to answer this part. If you want to look at a variety of lines of evidence, that would be wonderful but not necessary.

5. In class, I gave you formulae for the model averaged estimator for a parameter (e.g. \hat{N}) and the variance of that estimate (e.g. $\text{Var } \hat{N}$, incorporating model uncertainty). If you wanted a $1 - \alpha$ confidence interval for a parameter, you could use what I gave you to calculate $\hat{N} \pm z_{1-\alpha/2} \sqrt{\text{Var } \hat{N}}$. When we talked about specific models, I argued that profile likelihood confidence intervals are more appropriate than normal-based confidence intervals.

How might you calculate a profile-likelihood model-averaged confidence interval for a parameter? In other words, how might calculate a profile-likelihood confidence interval that accounts for model uncertainty?

Please include in your answer a list of issues and how your algorithm deals with them. For example, if I had asked about a “simple” profile likelihood interval, your list of issues might include:

Issue: The estimated population size is not normally distributed. The proposed method deals with this by basing the confidence interval on a likelihood-ratio hypothesis test that does not assume normality of the test statistic.

Notes: You don’t have to do the calculations. I only want you to describe an algorithm to calculate this.

You should know that I don’t know the answer! I don’t know whether anyone knows the answer. I don’t even know whether anyone has ever thought about this in a frequentist context. I know the answer in a Bayesian context, but we haven’t talked about Bayesian approaches.