

# Lecture 21

- **Logistic Regression Example (Critique)**  
Geist and Auerwald (2007)
- **Occupancy Methods (A Generalisation)**  
Methodology  
Examples
- **Survey Design Brief Critique**  
Brown and Harris (2005)

## Geist and Auerwald (2007): An Interesting Logistic Regression Example

- 26 streams in 7 European Countries. Streams were chosen based on previous or present occurrence of mussels. 15 streams had not had mussel recruitment in the last 30 years. Clearly not feasible to choose the streams randomly for many reasons.
- 275 Sites based on stretches of stream (5-50m lon) of which 46 Functional, 39 Potentially Functional and 190 Nonfunctional
- Many Measures taken on these sites.

# Geist and Auerwald (2007): An Interesting Logistic Regression Example

- We focus on the aspect of separating Functional and Nonfunctional Stream Sites 1-F, 0-NF. The analysis ignored the clustering in streams. (effect on SEs)
- Important Covariates geometric mean particle diameter ( $d_g$ ) and penetration resistance ( $r$ ).
- Logistic Equation

$$\text{Logit}(\pi_i) = -1.5 + 1.3d_g - 31.1r$$

- Mussels like coarser substrates and substrates that are not too compacted (ie lower penetration resistance).

# Geist and Auerwald (2007): An Interesting Logistic Regression Example

- Want to emphasise I am not over critical of this paper as this kind of large scale study is difficult to carry out.
- Clustering on sites ignored in the analysis (Discuss effects on SEs).
- Difficulties of choosing the sites to include in the survey for a rare and endangered species.
- Discuss what effect this has? I suspect they have included most streams that have good popns but what about the ones they included that are now nonfunctional I suspect they would have had thousands of more choices for those.
- Suppose someone who didn't know the current mussel distribution had tried to chose streams using a stratified random approach and then chosen sites nested within the streams with some minimal spatial separation.
- It sounds good but I suspect that almost all the sites would have had 0s!!!

# Sample Sizes in Logistic Regression

- I don't have good guidelines for logistic regression sample sizes. Roughly 50-100 points per parameter needed?

$$\hat{p} = \frac{x}{n}$$

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

If  $p = 0.1$  then  $n = 100$  gives  $SE = 0.03$  (30%)

If  $p = 0.1$  then  $n = 50$  gives  $SE = 0.042$  (42%)

If  $p = 0.5$  then  $n = 100$  gives  $SE = 0.05$  (10%)

If  $p = 0.5$  then  $n = 50$  gives  $SE = 0.071$  (14%)

- Previous example 3 parameters RSE approx 30% based on 236 points. (46 out of 236 sites or about 0.19 occupied)
- Shriner example –many parameters but she had about 2000 pts.

# Occupancy Estimation: Reference

Mackenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., and Hines, J. E. (2005).

Occupancy Estimation and Modeling :  
Inferring Patterns and Dynamics of  
Species Occurrence.

Elsevier, San Diego, USA.

Make Sure You all Run Out and Buy It So I  
get Some Royalties!

# Patch Occupancy Rate Metric: The Problem

- Conduct “presence-absence” (detection-non detection) surveys for a particular species of interest
- Estimate what fraction of sites (or area) is occupied by a species ( $\psi$ ) when the species is not always detected with certainty, even when present (i.e.  $p < 1$ ). We can also relate ( $\psi$ ) to important covariates as in regular logistic regression
- Naïve Occupancy rate estimates (Shriner et al paper) may be biased low because some sites where the species was not detected are occupied.
- In other words apparent absence of a species from a site may just be a failure to detect the species. This could be because individuals are hard to detect or because the species is very rare or both!!
- Also can we detect changes in the occupancy rate temporally or spatially.

# Model Parameters

$\psi_i$  -probability site  $i$  is occupied

$p_{ij}$  -probability of detecting the species in site  $i$  at time  $j$ , given species is present

Both parameters could be functions of covariates

# Data Summary: Detection Histories

- A detection history ( $h_i$ ) for each visited site or sample unit ( $i$ )
  - 1 denotes detection
  - 0 denotes nondetection
- Example detection history:  $h_i = 1\ 0\ 0\ 1$ 
  - Denotes 4 visits to site
  - Detection at visits 1 and 4

# Data Summary: Detection Histories: 4 visits to S sites. S is known.

Site	Detection History
------	-------------------

1	1101
---	------

2	1001
---	------

3	1100
---	------

4	0000-
---	-------

Note detections histories with all 0s allowed as part of data (Different to capture-recapture)

.

.

S	1011
---	------

# A Probabilistic Model: Very Similar to Capture-Recapture Models in Concepts Used

Sites that are occupied, For example

$\Pr(\text{detection history } 1001) = \Pr(h_i = 1001) =$

$$\psi_i [p_{i1}(1-p_{i2})(1-p_{i3})p_{i4}]$$

## Model: Key Issue, Apparent vs. True Absence

Sites where the species was not detected at all. These sites may or may not be occupied

$$\Pr(\text{detection history } 0000) = \Pr(h_k = 0000) =$$

$$\Psi_k \prod_{j=1}^4 (1 - p_{kj}) + (1 - \Psi_k)$$

**First Term-** Site is Occupied but Species Escapes Detection

**Second Term-** Site is Unoccupied

# Model: Covariates

- Site-specific: model  $\psi$  and/or  $p$ 
  - e.g., habitat type, patch size, patch isolation
- Survey-specific: model  $p$ 
  - e.g., local environmental conditions (water temperature)
- Use link function such as logistic:

$$p_{ij} = \frac{e^{\beta_0 + \beta_1 x_i + \beta_2 y_{ij}}}{1 + e^{\beta_0 + \beta_1 x_i + \beta_2 y_{ij}}}$$

# Model: The Likelihood Function

- The combination of these statements forms the model likelihood:

$$L(\underline{\psi}, \underline{p} | h_1, h_2, \dots, h_S) = \prod_{i=1}^S \Pr(h_i)$$

- Maximum likelihood estimates of parameters can be obtained
- Remember it may be more complex as covariates likely modelled
- Var-Cov matrix estimated using inverse of Fisher Information or parametric bootstrap
- Parameters cannot be site-specific without covariates.

# Model: Assumptions

- The detection process is independent at each site
- No heterogeneity that cannot be explained by covariates
- Sites are closed to changes in occupancy state between sampling occasions (That is sample in a short time window)

# Model: Software

- Windows-based software:
  - Program PRESENCE (J. Hines & D. MacKenzie) is available at the Patuxent Software Site
  - Program MARK (G. White)
- Fit both predefined and custom models, with or without covariates
- Provide maximum likelihood estimates of parameters and associated standard errors
- Assess model fit

# Another Example: Giant Weta Study

- An ancient giant insect (Order Orthoptera) now endangered in NZ due to introduction of small rodents. P116 in Occupancy Book.
- Therefore the need to look at occupancy within a reserve and how it is affected by habitat (browse vs non browse by cattle and goats)
- S=72 sites which are 3m radius plots searched between 3 and 5 times in March 2004

# Other Examples: Giant Weta Analysis

Occupancy covariate - browse or no browse

Detection probability covariates - day and  
observer

Best Model using AIC is  $\psi(\text{Browse})$ ,  $p(\text{day} + \text{observer})$

# Other Examples: Giant Weta Analysis

Occupancy Estimation

Naive  $\hat{\psi} = 0.49$  (SE = 0.06)

$\hat{\psi}(\text{Browse}) = 0.77$

$\hat{\psi}(\text{No Browse}) = 0.50$

$\hat{\psi} = 0.63$  (SE = 0.08)

Detection Probability Estimation

$\hat{p}$  varied widely from 0.10 - 0.69.

That is why the occupancy estimate of 0.63 is so much higher than the naive occupancy estimate of 0.49.

# Design Issues for Occupancy Studies

- What is a ‘season’?
- How to define a ‘sampling unit’?
- Selecting sampling units
- Repeat surveys
- More units vs. more surveys?

# What is a 'Season'?

- A season is a period of time during which it is reasonable to assume occupancy is static or changes occur completely at random
- Depends very much on the target species and study objective

# How to define a ‘sampling unit’?

- Should be assessed on a case-by-case basis
- Large enough to have a reasonable probability of occupancy, but not so large that any measure may be meaningless
- Size matters!

# How to define a ‘sampling unit’?

- Is there a natural definition? (Pond)
- At what scale do you want to measure occupancy?
- Is the species territorial?
- What density does it occur at?
- What is the size of it’s home range?

# Selecting sampling units

- How units are selected determines how results can be generalized
- Each sampling unit within the population should have a non-zero probability of being selected
- If units are selected such that occupancy is different than for the population of interest, estimates may be biased.
  - e.g., surveying only at historic sites

# Definition of Repeat Surveys

Repeat surveys do not (necessarily) imply repeat visits

- Discrete visits
- Multiple surveys within single visit
  - Single observer, conducting multiple surveys
  - Multiple observers each conducting a single survey
  - Multiple survey plots within a larger sampling unit

# Allocation of Effort

- For a ‘standard design’ there is an optimal number of repeat surveys per unit.
- The optimal number depends upon values for  $\psi$  and  $p$ .
- Does not depend upon number of units or total number of surveys.
- Reasonably robust to effect of cost.



## How many units?

- Once the number of repeat surveys has been determined, how many units to survey can be determined from the variance equation

$$\text{Var}(\hat{\psi}) = \frac{\psi}{U} \left[ (1 - \psi) + \frac{(1 - p^*)}{p^* - Kp(1 - p)^{K-1}} \right]$$

$$p^* = 1 - (1 - p)^K$$

## How many units?

- Example, if  $\psi \approx 0.7$  and  $p \approx 0.4$ , should use 5 surveys per unit;  $p^* = 0.92$
- To achieve a SE of 0.04 use  $S=183$  based on Equation 6.1 in the reference book and on the previous equation.

## General recommendations on allocating effort

- When detection probability is  $>0.5$ , at least 3 surveys per unit
- More surveys will be required when  $p$  is lower
- For rare species, survey more units less intensively
- Increasing spatial replication with insufficient repeat surveys may not be worthwhile

## General recommendations on allocating sampling effort

Example: if  $\psi \approx 0.4$  and  $p \approx 0.3$

Surveying 200 units twice gives  $SE(\hat{\psi}) = 0.11$

Surveying 80 units 5 times gives  $SE(\hat{\psi}) = 0.07$

- a decrease of 36%

With only 2 surveys per unit, would require 500 units to achieve same level of precision, or increase total effort 250%

# Non-standard designs?

- Repeatedly surveying a subset of units and elsewhere only once *does not* generally provide a more efficient design.
- Surveying a unit repeatedly until first detection (up to a maximum) may provide a more efficient design, but may be less robust. (Note the ‘optimal’ maximum number is higher than values given in the previous table)
- The standard design where each site visited the same no of times is probably best.

# Final Comments on Design

- Designing studies tends to be an iterative affair
- Simulation and pilot studies can provide useful information on how designs and field methods are likely to perform
- Program GENPRES (Hines) is available at the Patuxent Software Site.

# Multiple Seasons: Changes in Occupancy

- There are multiple visits within each of several seasons (Fig 7.1 from Mackenzie et al. 2005 attached)
- Within a season we have closure while between seasons there can be local extinction ( $\epsilon$ ) and also local recolonisation ( $\gamma$ ) (Fig 7.1 and 7.2 attached).
- This is similar in concept to the robust design used for capture-recapture (Ch 19 Text) and for looking at changes in communities (Ch 20 text)).
- This can form the basis for designing a long term monitoring study based on occupancy changes (amphibian surveys planned or implemented are using this approach).

# Conclusions

- A generalised logistic regression analysis which always for accounting for uncertain detection
- Tradeoff between cost and a robust model
  - There is a cost to replication
  - There is a large value to replication unless the detection probability is very close to 1.
- There is good software available to compute the estimates

# Brown and Harris(2005): Example of the Importance of Sample Survey Design

- The Case for citizen participation in the Adirondack (State Park NY) to Algonquin (Provincial Park, Ontario) corridor proposal.
- A survey of some of the resident landowners on the corridor path in NY State was an important part of the paper. (There are also private lands in Ontario as well).
- I thought this was an interesting paper even though it is about the design of a survey of people not animals or plants which has been our focus.
- Many of the same sampling principles we have discussed are very important when designing human surveys (often of wildlife or fishing use or about policy as here).
- There are also some unique features of human surveys which need to be considered.

# Brown and Harris(2005): Overall Results

- They managed to get 47 completed interviews which is 65% of those they attempted to get.
- 17% only knew something about the proposal. Despite weaknesses of the survey it is clear that citizens at the local level have not been involved to date in any meaningful way.
- They are Sceptical of economic benefits.
- They are Sceptical of and hostile to tourism in their area.
- Hunting was the most common land use activity the owners indulged in (64%). Even more than farming (43%).
- Very positive attitudes to protecting biological habitats despite their suspicion about the corridor project.
- Lets explore if we think the estimates are reasonable in terms of bias and precision.

# Brown and Harris(2005): Sampling Frame Issues

- The sampling frame here is a conceptual list of resident landowners in the corridor area in NY. ( $\{1,2,\dots,N\}$  where N is all the resident landowners as we discussed earlier under sampling section)
- There could have been definitional problems. They say that high density areas such as villages and town centers were avoided as well as houses on small lots.
- I suspect they considered only resident owners because this was a personal interview survey and they may have had a very hard time to reach nonresidents living in cities like Boston or NY City. (Phone Survey?).
- It is not clear to me that they actually had all the N in the frame when they started visiting people.

## Brown and Harris(2005): Non Response Issues

- They had a 65% response rate. It is not clear why the others did not respond. (Refusals vs. not at home are very different and it would have been good to have these enumerated separately)
- There is the potential for nonresponse bias. Let us discuss this for a while and the direction it probably goes in.

## Brown and Harris(2005): Response Issues

- There could be some biases induced by how the questions were asked. The questionnaire is not included.
- Question wording and interviewer effects could very important especially at this early stage where many landowners knew nothing of the corridor proposal.

# Brown and Harris(2005): Sample Size Issues

- The sample size was very very small especially as they then start comparing attitudes of large landowners (more than 100 acres) vs smaller landowners.

## Brown and Harris(2005): Final Comment

- The reason for the research very sound!
- The implementation of the research leaves a lot to be desired in my opinion.
- More like an extended op ed policy piece than true research study. Weird as the survey was not without cost.