Logistic Regression with Replicates

Some slides from Craig Burkett

STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

Krunnit Islands

- Natural wildlife reserve off Finnish coast of Gulf of Bothnia (north Baltic sea)

Study

- Want to know whether it is better to create reserves on larger or smaller islands
- Each island was visited in 1949 and 1959
  - If a species was found in 1949 but not in 1959, it was considered "extinct" (though possibly the researchers just didn't find any specimens…)
- (Data, in R)

Binomial Logistic Regression

- π_i - probability of extinction (in 1959)
- m_i - number of species (in 1949)
- y_i - number of species *gone* (in 1959)
- Assume species survival is independent, then
$$y_i \sim Binomial(m_i, \pi_i)$$
- Response proportion : $\hat{\pi}_{s,i} = y_i / m_i$
- Empirical (observed) logits: $\log\left(\frac{\hat{\pi}_{s,i}}{1-\hat{\pi}_{s,i}}\right)$

Binomial Logistic Regression

- (in R)

Ulkokrunni Island

- Predicted proportion

$$logit(\hat{\pi}_1) = -1.19 - 0.30\log(185.8) = -2.76$$

$$\hat{\pi}_1 = \frac{e^{-2.76}}{1 + e^{-2.76}} = 0.059$$

- Response proportion

$$\hat{\pi}_{s,1} = \frac{5}{75} = 0.067$$

Interpretation of coefficient

- For a log-transformed predictor
  $$logit(\pi) = \beta_0 + \beta_1 \log(x)\,, \beta_0 = -1.19, \beta_1 = -0.30$$
- Changing x by a factor of k changes odds by a factor of k^{β_1}
  - (Changes $\log(x)$ by $\log(k)$)
  - (Changes log-odds=$logit(\pi)$ by $\beta_1 \log(k)$)
  - (Changes odd by a factor of $\exp(\beta_1 \log(k)) = \exp\left(\log\left(k^{\beta_1}\right)\right) = k^{\beta_1}$)
- If you double the island area, odds change by a factor of k^{β_1}
- If you double the island area, odds change by:
  - $2^{-0.30} = 0.81$
- 95% CI:
  (in R)

Goodness of Fit

- Since we have several observations per “cell” (e.g., several observations for area), we can try to check whether the model fits the data
- “Saturated” model:
  - Each island has its own probability of extinction
- Our model:
  - The probability of extinction depends only on the log-area
- The saturated model fits the data better, but has fewer degrees of freedom (0)
- The model has (Npoints-Nparams-1) degrees of freedom
- The Null Model (fit a single probability to all islands) has

Goodness of Fit: Drop in Deviance

- aka Deviance goodness-of-fit test, compares:
  1. Model of Interest:
     $$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$
  2. Saturated model:
     $$logit(\pi_i) = \beta_0 + \beta_1 I_1 + \cdots + \beta_{n-1} I_{n-1}$$

H_0: Fitted model equivalent to saturated model

H_a: Saturated model better

- Test stat (G^2) will have a $\chi^2_{n-(p+1)}$ distribution under H_0, if all m_i are large (at least 5)

Drop in Deviance test

1. Model of Interest:

$$\hat{\pi}_{M,i} = \frac{e^{\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}}}$$

2. Saturated model:

$$\hat{\pi}_{s,i} = \frac{y_i}{m_i}$$

$$G^2 = -2(logL_M - logL_s) = 2(logL_s - logL_M)$$