STA 303 / 1002

Note Title 2/13/2012

TEST Monday Feb. 27

Where?

Last name	Location
A-K	EX 310
L-Z	EX 320

Bring: calculator + T-card

Coverage: Analysis of variance + logistic regression
To end of lecture on Thurs Feb 16

Formula page: similar to last page of last year's test

↑ I supply

Office hours during Reading Week (regular hours cancelled)

Thurs Feb 23 10:00 – 12:00 SS 5016A

Fri Feb 24 10:00 – 12:00

The Binomial Logistic Regression Model

Data: y_i observed Binomial counts, $i = 1, \ldots, n$

$$y_i \sim \text{Binomial}(m_i, \pi_i)$$

Model: $$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Example: Krunnit Islands – off Finland

Objective: Learn how to create nature preserves that help preserve endangered species.
Are large or small preserves better?

Have counts of bird species for 18 islands.

Data: - area of island in km^2
- number of species on each island in 1949
- " " " " " " in 1959

$n = 18$

π_i = probability of "extinction" (extinction = species no longer on particular island in 1959)
Assume that this is the same for each species of bird on a particular island

m_i = number of species on island i in 1949

y_i = number of species no longer there on island i in 1959

Assume species survival is independent

Then $y_i \sim \text{Binomial}(m_i, \pi_i)$

With these data (unlike Donner party example) can estimate π_i from data

"Response proportion" – est. of π_i from y_i, m_i

$$\hat{\pi}_{S,i} = \frac{y_i}{m_i} \qquad S = \text{"saturated"}$$

"Observed" or "Empirical" Logits: $\log\left(\frac{\hat{\pi}_{S,i}}{1-\hat{\pi}_{S,i}}\right)$

Proposed Model: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{area}$

Can plot observed logits versus area to see if a linear model seems appropriate

For our example, from this plot, decided we should look at log(area)

The relationship between log(area) and empirical logits seems linear

Model we will fit: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \log(\text{area})$

SAS: Model statement:

```
model yi / ni = xi ;
```

Need to specify both count of # of successes and number of trials

Output:

Number of observations: 18 (n) (# of islands)

Sum of frequencies: 632 $= \sum_{i=1}^{18} m_i$

Fitted model: $\log i(\hat{\pi}) = -1.196 - .297 \log(\text{area})$

Wald test: strong evidence that coef of log(area) is not zero ($p < .0001$)

Same test as in binary logistic regression

$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

Test stat $z_{obs} = \frac{-.2971}{.0549}$ or $\left(\frac{-.2971}{.0549}\right)^2 = 29.3$

Chi-square (1)

95% CI for β_1: $-.2971 \pm 1.96\,(.0549)$

$= (-1.19, -.04)$

Interpreting β_1 (since area was log-transformed)

$$\text{logit}(\pi) = \beta_0 + \beta_1 \log(x)$$

$$\Rightarrow \frac{\pi}{1-\pi} = e^{\beta_0} e^{\beta_1 \log(x)}$$

Changing x by a factor of k, changes odds by a factor of k^{β_1}

Krunnit example: Double island area $\beta_1 = -.2071$

odds change by a factor of $2^{-.2071} = .81$

So odds of extinction are 81% of odds on larger island

island half its size.

Likelihood Ratio Tests are same as for

binary logistic regression:

Fitted model: $\text{logit}(\hat{\pi}) = -1.196 - .297 \log(\text{area})$

For Ulkokrunni ($i = 1$), area $= 185.5$ km^2

$\text{logit}(\hat{\pi}_1) = -1.196 - .297 \log(185.5)$

$= -2.75$

Estimated prob. of extinction for a species on Ulkokrunni

$$\hat{\pi}_1 = \frac{e^{-2.75}}{1 + e^{-2.75}} = 0.060$$

Call this $\hat{\pi}_{M,1}$, "M" is "for model"

For comparison, the response proportion for Ulkotrunni is $\hat{\pi}_{S,1} = \frac{5}{75} = .067$

EffectPlot.png (add plots(only) = effect on proc logistic statement)

- plots $\hat{\pi}_M$ & $\hat{\pi}_S$

↑ with smooth curve

↑ points

What's the big difference between binary and binomial logistic regression?

Can do more tests for model adequacy in binomial LR.

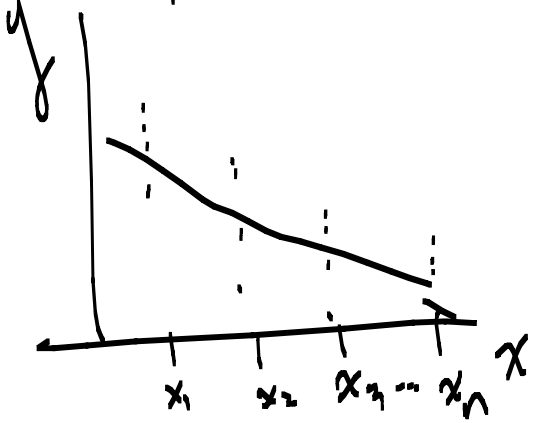
One such test: DEVIANCE GOODNESS-OF-FIT TEST (GOF)

Is a linear model appropriate / adequate?

The idea behind this test

(explained in terms of simple regression)

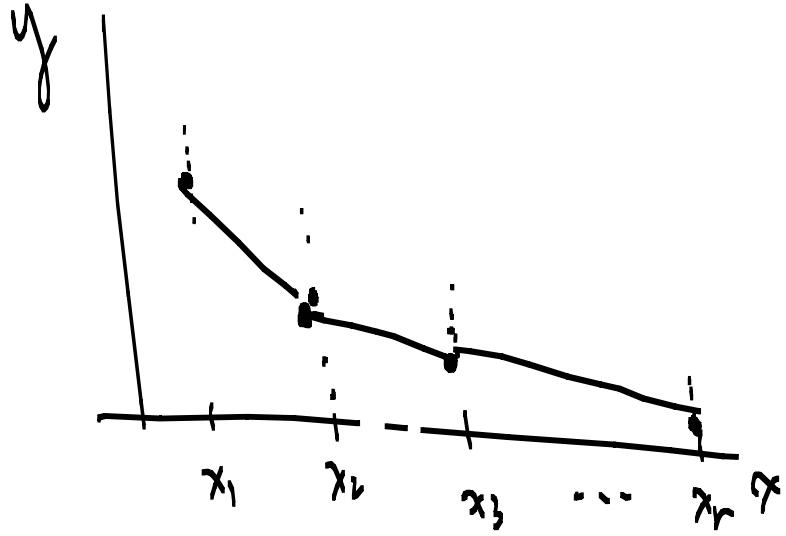
Suppose we have Replicate observations at each value of x (n different values of x)



← $y = \beta_0 + \beta_1 x + \varepsilon$

Alternative model:

Analysis of variance: $y = \beta_0 + \beta_1 I_{x=x_1} + \beta_2 I_{x=x_2} + \cdots + \beta_{n-1} I_{x=x_{n-1}} + e$



- more β's but better model if relationship isn't linear.

There is a linear regression GOF test that compares these 2 models

In logistic regression for binomial counts, think of

data for each x_i as m_i Bernoulli observations.;
then have m_i observations at x_i

The Deviance G-O-F test compares:

① Model of Interest: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$ (2 parameters)

to ② Saturated Model: $\text{logit}(\pi) = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \ldots \beta_{n-1} I_{n-1}$ (n parameters)

In saturated model, each explanatory variable is an indicator variable; $x^{(i)}$ treated as a categorical variable with n categories

Use Likelihood Ratio Test to compare Model of Interest (reduced model) to Saturated Model (full model)

This called the "Drop-in-Deviance" test (sometimes)

Test statistic is $-2\log\left(\frac{L_R}{L_F}\right) = -2\log\left(\frac{L_M}{L_S}\right)$

H_0: fitted model fits data as well as saturated model

H_a: saturated model is better

Under H_0, the test statistic is an observation from a chisquare distribution with $n-(p+1)$ df

(This is an asymptotic approximation; works well if each $m_i > 5$)

Calculation of test statistic:

For saturated model (full model) $\hat{\pi}_{s,i} = \frac{y_i}{m_i}$

For fitted model (reduced model) $\hat{\pi}_{M,i}$, estimated using MLE

$\hat{y}_i = m_i \hat{\pi}_{M,i}$

For binomial logistic regression,

Likelihood function

$$L = \prod_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}$$

where $\pi_i = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}$

$$\log L = \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (m_i - y_i) \log(1-\pi_i) + \log \binom{m_i}{y_i} \right]$$

Test statistic for Deviance GOF test is called the "deviance"

$$\text{Deviance} = -2\left(\log L_M - \log L_S\right) = 2\left(\log L_S - \log L_M\right)$$

$$= 2 \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{m_i}\right) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i}\right) + \log \binom{m_i}{y_i} \right.$$
$$\left. - y_i \log\left(\frac{\hat{y}_i}{m_i}\right) - (m_i - y_i) \log\left(\frac{m_i - \hat{y}_i}{m_i}\right) - \log \binom{m_i}{y_i} \right\}$$

$$= 2 \sum_{i=1}^{n} \left\{ y_i \log(y_i) + (m_i - y_i) \log(m_i - y_i) - y_i \log(\hat{y}_i) - (m_i - y_i) \log(m_i - \hat{y}_i) \right\}$$

Small deviance $\Rightarrow$ model fits data well

To get deviance GOF test in SAS, use scale = none option on model statement

Krunnit example: Test statistic = 12.06

H_0: fitted model is equal in fit to saturated model

H_a: saturated model fits better

$df = 16 \; (= 18 - 2)$

$p = .7397$

The data are consistent with H_0; we'll use the simpler model with (linear function of log(area))

For the deviance GOF test:

- Large p-value means
  - fitted model is adequate
  - OR - test is not powerful enough to detect inadequacies
- Small p-value means
  - the fitted model is not correct (eg. need more explanatory variables, or need polynomial model, or ...)
  - OR - response distribution is not adequately modelled

by the Binomial distribution

OR - there are some severe outliers