STA 303 / 1002

Note Title 2/6/2012

Reminder: Assignment 1 due Thurs Feb 9 at 2:00 pm.
(don't be late.)

Logistic Regression Model

Response Y_i, $i = 1, \ldots, n$ is a Bernoulli r.v.
with probability of success is π_i
where π_i depends on values of p explanatory variables
$X_{i1}, \ldots, X_{ip}$

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

Use MLE to find estimates of the β's

Donner Party Example

outcome — DIED
\ SURVIVED

, $n = 45$, "adults" $\geq$ 15 years old.

predictors — age
— sex — FEMALE
\ MALE

New SAS procedure: proc logistic

By default

① - will model P(DIED) (DIED comes before SURVIVED alphabetically)

- use descending option to change this

② Class statement for categorical predictor variables uses 'effect coding' (1/-1) rather than indicator variables (1/0)

- use /param = ref; to change to indicator variables

- model will 1 = female
0 = male

Fitted equation for defaults:

π = P(DIED) fitted equation: $\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.43 + .078\,\text{age} - 0.80\,\text{sex}$

Using my preferences:

π = P(SURVIVED)

fitted equation: $\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 1.63 - .078\,\text{age} + 1.60\,I_{\text{female}}$

(Exercise: Convince yourself that these 2 fitted equations agree.)

So for 25-year-old female:

estimated odds of survival: $\frac{\hat{\pi}}{1-\hat{\pi}} = \exp\{1.63 - .078(25) + 1.60(1)\}$

Estimated probability of survival: $\hat{\pi} = \dfrac{\exp(\overbrace{1.28}^{1.28})}{1+\exp(1.28)}$

$= 0.78$

So 25-year-old female had 0.78 prob. of survival

Interpreting Coefficients in Logistic Regression

Let w be the odds that $Y=1$; $w = \exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}$

Interpretation of β_1: hold $X_2, \ldots, X_p$ fixed

Then the ratio of the odds that $Y=1$ at $X_1 = a$ to $X_1 = b$.

$$\frac{w_a}{w_b} = \exp\{\beta_1(a-b)\}$$ "ODDS RATIO"

If X_1 increases by 1 unit, all other X's held constant, the odds that $Y=1$ change by a multiplicative factor e^{β_1}

Donner Party Example Fitted model: $\text{logit}(\hat{\pi}) = 1.63 - .078\,\text{age} + 1.6\,I_{\text{female}}$

Eg. Compare 50 year old woman to 20 year old woman
Estimated odds ratio of survival $= \exp\{-.078(50-20)\}$
$= .096 \sim \frac{1}{10}$

Odds of survival for 20-year-old are 10 times

the odds for a 50-year-old.

- Compare woman to man of same age:
  Estimated odds ratio = $\exp\{1.6(1-0)\} = 4.95$

  So a woman's odds of survival are 5 times those of a man of the same age.

Tests for whether β's are different from 0:

(if $\beta_1 = 0$, X_1 has no effect on (log-odds))

Based on large-sample properties of MLEs

As $n \to \infty$, MLEs are normally distributed

Wald test: $H_0: \beta_j = 0$ vs $H_a: \beta_j \neq 0$

Wald test statistic: $z_{obs} = \dfrac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$

z_{obs}

↳ estimated from numerical procedure that generates MLEs

For large n, if H_0 is true, z_{obs} is approx. an observation from a standard normal distribution

100$(1-\alpha)$% CI for β_j (Wald): $\hat{\beta}_j \pm z_{\alpha/2}\, s.e.(\hat{\beta}_j)$

$z_{\alpha/2}$ ← $\alpha/2$ quantile from $N(0,1)$

Corresponding SAS output

Parameter	df	Estimate	S.E.	Wald Chi-square	Pr > Chisq
SEX FEMALE	1	1.60	.756	4.47	.0345

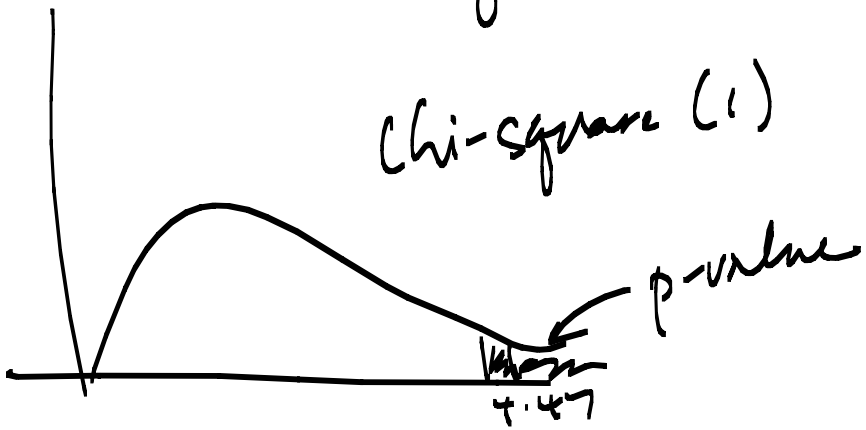
$$\left(\frac{1.60}{.756}\right)^2 = \left(\frac{\text{Est.}}{\text{S.E. of Est.}}\right)^2$$

Z_{obs}

P-value

$N(0,1)$

Chi-square (1)



(Recall: square of a std. N r.v. has a Chi-square distribution with (df))

Conclude: Moderate evidence that both age ($p = .036$) and sex ($p = .0345$) have an effect on odds of survival over and above each other

95% CI for coefficient of age: $-.078 \pm 1.96(.756)$
$= (-.15, -.0055)$

For odds ratio for 1 year increase change in age, same sex

CI: $(e^{-.15}, e^{-.0055}) = (.86, .99)$

Odds of survival for 50-year-old vs 25-year-old (same sex):

log odds change by $-.078(25) = -1.95$

95% CI for change in log odds $25 \times (-.15, -.0035)$

$= (-3.76, -.137)$

Odds of survival for a 50-year old are $e^{-1.95} = .14$ times the odds of survival for a 25-year old

95% CI for odds ratio $(.023, .87)$

Not appropriate to calculate a CI for π because $0 \leq \pi \leq 1$, so <u>Not</u> normally distributed

Likelihood Ratio Tests (LRTs)

Idea: comparison of full and reduced models.

reduced: missing 1 or more predictor variables from full model

- same data

Compare likelihood of data assuming full model (L_F) to likelihood assuming reduced model (L_R)

Likelihood ratio: $\frac{L_R}{L_F}$ $\quad L_R \leq L_F$ (get larger maximum with bigger model)

H_0: reduced model is appropriate (fits data as well as full model)
(extra parameters that are in full model are 0)

H_a: full model is better

Test statistic $G^2 = -2\log\left(\frac{L_R}{L_F}\right)$

"log" = NATURAL LOG = "ln" always

For large n, if H_0 is true,
G^2 is an observation from a chi-square distribution with
df = difference in number of parameters between full and reduced models

Note For testing whether 1 parameter is 0, could use either Wald test or LRT

They aren't equivalent

If they don't agree, use LRT. It has been shown to be more reliable.

SAS gives "global" LRT

- compares fitted model to null model

$\log(\pi/(1-\pi)) = \beta_0$

Test: $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

vs H_a: at least one of $\beta_1, \ldots, \beta_p$ is not 0

(SAS also gives Score and Wald test for same H_0 and H_a which we won't cover)

Donner Example Model $\text{logit}(\pi) = \beta_0 + \beta_1 \text{age} + \beta_2 I_{\text{female}}$

Global LRT: $H_0: \beta_1 = \beta_2 = 0$ vs H_a: at least one of β_1, β_2 not 0

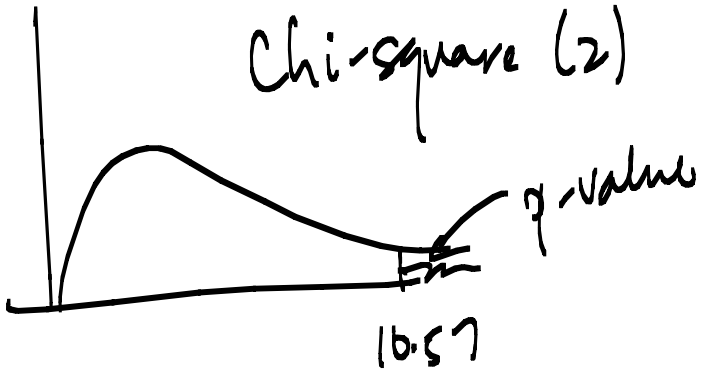
Test stat: $G^2 = -2 \log\left(\frac{L_R}{L_F}\right)$

$= -2\log(L_R) - (-2\log(L_F))$

$= 61.827 - 51.256$

$= 10.57$

If H_0 true, this is an observation from Chi-square (2) distribution



$p = .0051$

Strong evidence that fitted model is better than the null model

At least one of age, sex is useful for explaining odds of survival.

Model Assumptions: