STA 303 / 1002

Note Title 2/9/2012

Model Assumptions for Logistic Regression

- fewer assumptions than in ordinary linear regression

  (Not true of logistic regression – const var.
  - $E(e) = 0$ (no error terms)
  - normality)

For logistic regression, model assumptions:

- independent observations
- Correct form of model

- linearity between logits and explanatory variables
- include all relevant variables
- exclude all irrelevant variables
    - don't want to have multicollinearity

- need large samples for tests and CIs to be valid.
  (rely on large sample properties of MLEs)

For Donner Party Example, there were families in parties,
so we have concerns about independence of observations

<u>Checking model adequacy</u>

- no residual plots - no need to check for outliers
  (Y is only 0 or 1)

or const var. (not true here.

- residuals don't have a meaning in binary logistic regression

<u>Can</u> do: - check higher order polynomial terms and/or interaction terms improve model

Try <u>adding</u>: age * sex interaction,

age^2 non-linear term

age^2 * sex interaction

<u>Model</u>: $\text{logit}(\pi) = \beta_0 + \beta_1 \, age + \beta_2 I_F + \beta_3 \, age * I_F + \beta_4 \, age^2 + \beta_5 \, age^2 * I_F$

Is this model better than the model with just age & sex?

Are the coefficients of the extra terms 0?

$H_0: \beta_3 = \beta_4 = \beta_5 = 0$ vs H_a: at least one of $\beta_3, \beta_4, \beta_5$ not-zero

Likelihood Ratio Test

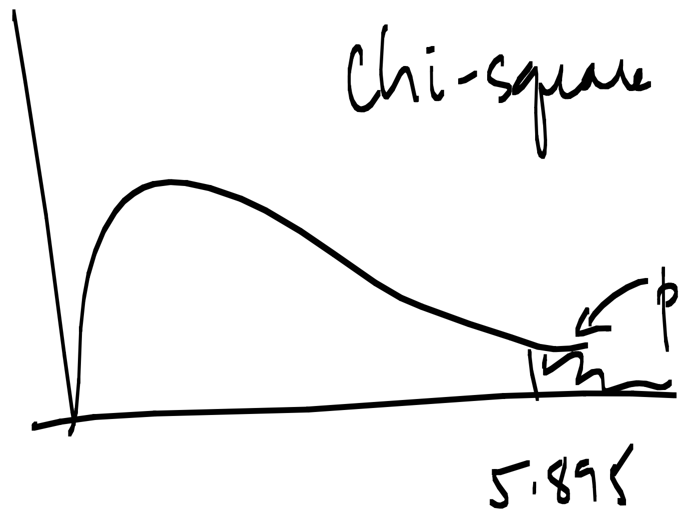
Test statistic: $G^2 = -2\log L_R - (-2\log L_F)$

$$= 51.256 - 45.361$$

$$= 5.895$$

Under H_0 this is an observation from a chi-square distribution with $\underline{3}$ df

Chi-square (3)



$p = 0.117$

So data are consistent with $\beta_3, \beta_4, \beta_5$ all 0

So go with original model

Predictors:
age - continuous var
sex - categorical

Additive Model: $\text{Logit}(\pi) = \beta_0 + \beta_1 \, age + \beta_2 \, sex$

"Proportional Odds" model

But seems reasonable that the effect of age depends on sex.

Consider model

$$\text{logit}(\pi) = \beta_0 + \beta_1 \, age + \beta_2 I_F + \beta_3 \, age * I_F$$

$H_0: \beta_3 = 0$ vs $H_a: \beta_3 \neq 0$

Wald test: p-value = .0865

Weak evidence that the effect of age on odds of survival differs with sex

Since only weak evidence, check with LRT (Exercise)

(A: $p = .048$)

Review of Binomial Distribution

If $Y \sim \text{Binomial}(m, \pi)$

$$P(Y = y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \quad y = 0, 1, \ldots, m$$

$Y = \sum_{i=1}^{m} X_i$ where X_i are independent Bernoulli (π) r.v.'s

$$E(Y) = m\pi$$
$$Var(Y) = m\pi(1-\pi)$$

Often consider Y/m = proportion of 1's

$$E(Y/m) = \pi$$

$$\mathrm{Var}(Y/m) = \frac{\pi(1-\pi)}{m}$$