STA 303 / 1002

Note Title 3/8/2012

Poisson Regression continued:

Elephant example: Is there a relationship between and and # of matings?

Model: $\log(\mu_i) = \beta_0 + \beta_1 \, age_i$

where μ_i is mean # of matings for elephant with $age = age_i$

Underlying prob. distn: counts of # of matings for an elephant with $age = age_i$ is Poisson with mean μ_i

SAS <u>output</u>

<u>Deviance</u>

Test stat for Deviance goodness-of-fit

H_0: fitted model (see above) fits as well as saturated model

H_a: saturated model fits better

<u>Saturated model</u>: introduces dummy variables for each value of age

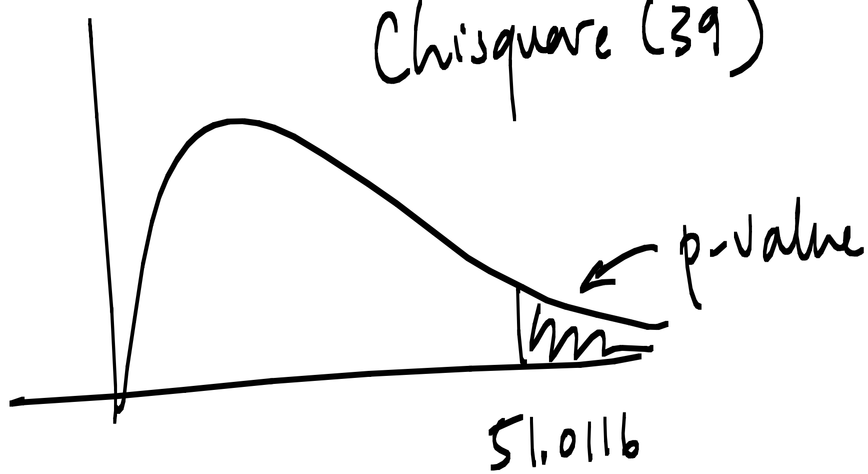
\# of parameters: fitted model is 2

saturated model is 41 (β_0 + β_i's for 40 indicator variables since $n = 41$)

# of df = 41 - 2 = 39

Test statistic: $G^2 = -2\log\left(\frac{L_R}{L_F}\right) = 51.0116$

Chisquare (39)



For a chisquare distribution mean $\approx$ df

SAS does not give the p-value

From R: $p = .0943$

Extremely weak evidence that saturated model fits the data better than the linear model

<u>Summary of findings</u>

Fitted equation: $\widehat{\log(\mu)} = -1.582 + .0687\,\text{age}$

Strong evidence that mean number of matings depends on age ($p < .0001$ (Wald test))

Estimated mean number of matings increases by a factor of $e^{.0687} = 1.071$ (7% increase) for a 1 year increase in age

Example : Framingham Heart Study

- town in Massachusetts. In 1948, recruited 5209 healthy men and women, aged 30-60 and followed (and still following their descendants) to examine risk factors for cardio-vascular disease (CVD)

Data we'll consider: 1329 men

- Chol. measured in 1948
- did they develop CVD in 10 years?

(mg/dL)	CVD		
Cholesterol	present	absent	total
≥ 260	41	245	286
< 260	51	992	1043
	92	1237	1329

Is high cholesterol associated with increased risk of CVD?

<u>Analysis 1</u> Let π_H be prob. of CVD in men with high cholesterol

Let π_L " " " " " "low" cholesterol

$H_0: \pi_H = \pi_L$

$H_a: \pi_H \neq \pi_L$

Est of π_H: $\hat{\pi}_H = 41/286$

" " π_L: $\hat{\pi}_L = 51/1043$

<u>Test statistic</u>: $\dfrac{\hat{\pi}_H - \hat{\pi}_L}{\text{s.e. of } (\hat{\pi}_H - \hat{\pi}_L)}$

Each person is a Bernoulli trial with chance of developing CVD $= \begin{cases} \pi_H \\ \pi_L \end{cases}$ $\quad n_H = 286$, $n_L = 1043$

Then (as long as people are independent), the count of the number of people who developed CVD is

$$\nearrow \text{Binomial}(286, \pi_H)$$

or

$$\searrow \text{Binomial}(1043, \pi_L)$$

$$\text{Var}(\hat{\pi}_H) = \frac{n_H \pi_H (1-\pi_H)}{n_H^2} = \frac{\pi_H (1-\pi_H)}{n_H}$$

more to come....