STA 303 | 1002

Note Title 3/1/2012

A few more things about Logistic Regression

More Model Fit Statistics

- useful for comparing models with same response and same data

- log-likelihood with penalty for number of parameters in the model

AIC - Akaike's Information Criterion

$$= -2 \log L + 2(p+1)$$

where p is the number of explanatory variables (so $p+1$ β's in the model)

- smallest is best

SC - Schwarz's (Bayesian) Criterion

$= -2\log L + (p+1)\log N$, where $N = \sum_{i=1}^{n} m_i$

- smallest is best
- larger penalty on model complexity than AIC

Rule of Thumb for AIC

- one model fits data better than another if difference in AIC's is > 10

- one model is essentially equivalent to another if the difference in AIC's < 2

Donner Party Example

First: (1) fit model with age and sex

+ More complicated models:

- (2) with age, sex, age^2, age*sex, age^2*sex
- (3) with age, sex, age*sex

Model	AIC
(1)	57.256
(2)	57.361

3 55,346

Choose Model 1 because ~~all~~ Δ AIC between 1 & 3 are within 2 of each other ; choose simplest model.

~~(2 is definitely worse than 3)~~

Some indication that 2 is worse than 3

Problems and Complications that can occur in linear regression that can also occur in logistic regression

- Extrapolation - don't make predictions / inferences

outside range of observed data ;
model may no longer be appropriate

- <u>Multicollinearity</u> – dependency among explanatory variables
  - consequences in linear regression also apply to logistic regression
- <u>Influential Points</u> – an observation is influential if its removal substantially changes estimated coefficients
- <u>Model Building</u> – choosing predictor variables tends to overfit the data
  - should build on training data

and test on test data

Two problems that can happen in logistic regression that weren't a concern in linear regression

① Extra-binomial variation
- variance of Y_i greater than $m_i \pi_i (1-\pi_i)$
- also called "overdispersion"
- doesn't bias estimates of β's but s.e. of $\hat{\beta}$'s will be estimated as smaller than they should be

What you can do: - fit one more parameter

(dispersion parameter, ψ)

$Var(Y_i) = \psi m_i \pi_i (1-\pi_i)$

(2) Complete and Quasi-complete separation

(see practice problem about Swiss banknotes)

Complete separation - linear function of explanatory variables perfectly divides the observations by whether $Y=1$ or $Y=0$

What happens: can't compute MLEs

Quasi-complete separation - explanatory variables predict $Y=1$ or $Y=0$

almost perfectly (just a few points wrong)

- MLE is numerically unstable

What to do: - simplify the model

Other options (beyond scope ...)
penalized max. likelihood, exact logistic regression

Not responsible for on logistic regression output

- Numbers under "Association of Predicted Probabilities and Observed Responses"

- For Global null hypothesis: Score and Wald tests
- Pearson G-O-F test