Generalized Linear Models(Lecture 5)

Zhengfan Wang

September 26, 2017

Variance/Covariance “wrap up”

$\beta^{(t+1)} = \beta^{(t)} + (I^{(t)})^{-1} * \mu^{(t)}$,

where I is variance/covariance matrix of estimate using Fischer scoring.

– Characterizing uncertainty

– Approximation to the likelihood or posterior surface

– Frequentist Method/Fisher Scoring

- relying on large sample approximation to likelihood surface
- relies on distributional assumptions

– Other options

- more computationly intensive
- hard to use w/black box
- sampling based (Bayesian MCMC;Gibbs sampling)

Inference Miscellany

Setting:logistic regression and the model is:

$$logit(\pi_i) = \alpha + \beta x_i$$

In large sample and $x = x_k$, the SE for

$$logit(\hat{\pi_k})$$

is described as:

$$\sqrt{var(\hat{\alpha} + \hat{\beta} x_k)}$$

.

And the variance can be written as

$$var(\hat{\alpha} + \hat{\beta} x_k) = var(\hat{\alpha}) + {x_k}^2 var(\hat{\beta}) + 2 x_k cov(\hat{\alpha}, \hat{\beta})$$

- Confidence Interval

`95% C.I. for`

$$logit(\hat{\pi_k}) = logit(\hat{\pi_k}) \pm 2 * SE(logit(\hat{\pi_k}))$$

```
If the C.I. is (a,b), 95% C.I. for
```

$$\hat{\pi_k} \in (\frac{e^a}{1+e^a}, \frac{e^b}{1+e^b})$$

```
R function is
```

predict(mymodel,type="response")

Model Checking and Building

Checking: does the model fit data well?

- residual plots/analysis
- predicted v.s. observed
- instability in model estimate (large SEs?) colinearity?

Building: process by which you arrive at a "chosen" model

- systematic/pre-specificed analysis plan

– corvariate choices

– selection criteria. e.g. AIC(Akaike information criterion),BIC(Bayesian information criterion), Likelihood-ratio test

– describe/explain analysis

– specify type 1 error

– specify correction for multiple testing

– specify validation SE

- iterative/subject

– use residual plots or other G.O.F measure to modify model

– incorporate domain-specific knowledge, e.g. other corvariates

– vigilant about "gorden of forking paths"

– important to use validation samples

– penalized regression (LASSO(least absolute shrinkage and selection operator, often used in Bioinformatics), GAMs,. . .)

– ensemble of "reasonable" model

Poisson GLMs

$$y_i \sim poisson(\lambda_i)$$

$$\eta_i = X_i\beta$$

$$g(E(y_i)) = log(\lambda_i) = \eta_i = X_i\beta$$

Loglink

- implies covariates have multiplicate effect.

$$\lambda_i = e^{\beta_0} * e^{\beta_1 x_1} * e^{\beta_2 x_2} * ...$$

- relative risk/rate interpretation for e^{β_k}

$$log\lambda_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

$$(log\lambda_i \mid x_1 = k+1, x_2 = j) = \beta_0 + \beta_1(k+1) + \beta_2 j$$

$$(log\lambda_i \mid x_1 = k, x_2 = j) = \beta_0 + \beta_1 k + \beta_2 j$$

$$\beta_1 = (log\lambda_i \mid x_1 = k+1, x_2 = j) - (log\lambda_i \mid x_1 = k, x_2 = j) = log(\frac{\lambda_i \mid x_1 = k+1}{\lambda_i \mid x_1 = k}) = log(RR)$$

$$RelativeRisk = e^{\beta_1}$$

Holding all other variables constant, the expected value of y change $[(e^{\beta} - 1) * 100]\%$, e.g.$(0.8 - 1) * 100 \rightarrow -20\%$ decrease; $(1.2 - 1) * 100 \rightarrow 20\%$ increase.

Exposure/Offset

outcome of interest e.g.1 rank of hospitalizations by county

$y_i = \#$ of times of outcome occured

$u_i =$ offset or exposure

e.g.2

$y_i = \#$ of case of flu in a population

$u_i =$ Population

e.g.3

$y_i = \#$ of traffic accident at intersection i in one day

$u_i =$ 1.average number of vichicles at intersection today. 2.average number of vichicles at intersection yesterday

$$y_i \sim poisson(\mu_i \lambda_i)$$

$$E(y_i) = \mu_i \lambda_i$$

$$log(E(y_i)) = log\mu_i + log\lambda_i = log\mu_i + X\beta$$

which $log\mu_i$ is the offset and $X\beta$ is the linear predictor.

$$log(\lambda_i) = \eta_i = x_i\beta$$

And in R

$glm(y \sim x_1 + x_2, offset = log(a), family = poisson, ...)$ offset is not aleardy on log scale.