2/03/16 ALSM 2

INFERENCE

Algorithms To calculate β's

Usually $\hat{\theta} = \arg\max_{\theta} \log[\text{like}(\theta)]$ usually differentiable this function and equate to 0

Normal Distribution

link y: identity function [mapping R onto R]

$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(y-\theta)^2}{2\sigma^2}\right]$ $\theta_i = x_i^T\beta$ - combining of mean.

likelihood wrt some parameter β: $\text{lik}(\beta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(y_i - \theta_i)^2}{2\sigma^2}\right]$

$\log(\text{lik}(\beta)) = -N\log(\sqrt{2\pi}\sigma) - \sum (y_i - x_i^T\beta)/2\sigma^2$

$SSE = \sum (y_i - x_i^T\beta)^2$

$\hat{\beta} = \frac{d(SSE)}{d(\beta)} = 0$ (how we get our βs)

$SSE = (\|\vec{y} - X\beta\|)^2$ (norm)^2

$\vec{y} = \begin{bmatrix} y_1 \\ y_n \end{bmatrix}$ $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{n} \\ \vdots & & & \\ 1 & x_{1N} & \cdots & x_{pN} \end{bmatrix}$

When we compute derivative of SSE: $\frac{d(SSE)}{d(\beta)} = \frac{d[(\vec{y} - X\beta)^T(\vec{y} - X\beta)]}{d(\beta)}$

$= \frac{d}{d\beta}[\vec{y}^T\vec{y} - \vec{y}^T X\beta - (X\beta)^T\vec{y} + (X\beta)^T X\beta]$

$= 0 + (\vec{y}^T X)^T - X^T\vec{y} + 2X^T X\beta = 0$

$\hat{\beta} = (X^T X)^{-1} X^T \vec{y}$

1. Differentiate lg likelihood $\nabla \log lik$
2. Set $\nabla \log lik = 0$
3. Solving system of equations defined by $\nabla \log lik = 0$

Newton-Raphson

- looking at $\log(\text{like}(\theta))$ near θ'

$\log lik(\theta) = \log lik(\theta^i) + (\theta-\theta^i)^T \nabla \log lik(\theta^i) + \frac{1}{2}(\theta-\theta^i)[\nabla\nabla^T \log lik(\theta^i)](\theta-\theta^i) + \text{remainder}$

$\nabla \log lik(\theta^i)$: gradient at θ^i

$\nabla\nabla^T \log lik(\theta^i)$: Hessian matrix at θ^i : $H_{(i)}$

After differentiation $\nabla \log lik \simeq H_{(i)}(\theta-\theta^i) + \nabla \log lik(\theta^i) = 0$

Newton-Raphson Method

Introduce $\theta^i = \theta^0$

$\theta^{(i+1)} = \theta^{(i)} - [H_{(i)}]^{-1} \nabla \log lik(\theta_i)$ (inverse * gradient)

Until Convergence

- If loglik is quadratic, NR will find solution in one step (i.e. normal distribution)
- In GLM, only one extrema, no local extrema
- The quantity $[-H]$ (Hessian - observed information) determines the sharpness of the peak in the likelihood function and its maximum

Alternative to Newton-Raphson: Method of Scoring

$I = E[J|\theta]$ (Expected Fisher Information)

Replace J by I in NR method

Multinomial Distribution and Poisson Random Variables

- Consider 2 independent random variables that follow a poisson distribution, i.e.

$y_1 \sim P_o(\lambda_1)$ Show that $n = y_1 + y_2 \sim P_o(\lambda_1+\lambda_2)$

$y_2 \sim P_o(\lambda_2)$

$P(y|\lambda_i) = \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i)$

$P(n) = \sum_{y_2=0}^{n} \sum_{y_1=0}^{n} P(n, y_1, y_2)$ Partition law

$= \sum \sum P(n|y_1,y_2) P(y_1,y_2)$ (Bayes) (independent)

$= \sum \sum \delta(n-(y_1+y_2)) P(y_1) P(y_2)$

10/3/16 AUSM2

Integrate: $= \sum_{y_2=0}^{n} P_{y_1}(n-y_2) P_{y_2}(y_2)$

$= \sum_{y_2=0}^{n} \frac{\lambda_1^{(n-y_2)}}{(n-y_2)!} \exp(-\lambda_1) \frac{\lambda_2^{y_2}}{y_2!} \exp(-\lambda_2)$

$= \sum_{y_2=0}^{n} \frac{\lambda_1^{n-y_2}}{(n-y_2)!} \frac{\lambda_2^{y_2}}{y_2!} \exp(-\lambda_1-\lambda_2)$ expecting: $P(n) = \frac{(\lambda_1+\lambda_2)^n}{n!} \exp(-(\lambda_1+\lambda_2))$

$= \sum_{y_2=0}^{n} \frac{n!}{y_2!(n-y_2)!} \lambda^{n-y_2} \lambda_1^{y_2}$

Show

Joint probability $p(y_1, y_5 | n) = \frac{n!}{y_1! \cdots y_5!} \theta_1^{y_1} \cdots \theta_5^{y_5}$

Given $y_i \sim P(\lambda) \quad \forall i = 1, 5 \quad n = y_1 + y_2 + \cdots + y_5$ with $\theta_i = \frac{\lambda_i}{\sum_{i=1}^{5} \lambda_i}$

$p(y_1, \ldots, y_5 | n) = \frac{p(n | y_1, \ldots, y_5) \, p(y_1, \ldots, y_5)}{p(n)}$ Bayes

$p(y_1, \ldots, y_5) = \prod_{i=1}^{5} \text{Poisson}(y_i | y_5)$

$p(n) = \text{Poisson}(n | \lambda_1 + \lambda_2 + \cdots + \lambda_5)$

$= \frac{\delta(n - (y_1 + \cdots y_5)) \prod_{i=1}^{5} \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i)}{\frac{[\sum_{i=1}^{5} (\lambda_i)]^n \exp(-(\lambda_1 + \cdots + \lambda_5))}{n!}}$

Limitation of likelihood function

- Assume responses are independent
- Outliers in data can give a zero, all data except one group votes for parameter β
- May not be robust because of outliers
- Hard to remove outliers in large datasets
- Some problem with bayesian approach

Linear Regression Scott Paper

$p(y | x \beta) = N(y, x^T \beta, \sigma^2)$ Hypothesis / Model

$\{ \varepsilon_i = y_i - x^T \beta \}$ $\varepsilon = y - x^T \beta \sim N(\varepsilon, 0, \sigma^2)$

↓ take empirical distribution

$p(\varepsilon) = \frac{1}{N} \sum_{i=1}^{N} \delta(\varepsilon, \varepsilon_i)$

↑ Dirac function

Get empirical distribution as close as possible to model distribution
i.e. get β for a minimising distance between two functions

Scott: Sum (not product): L.E $\{ \Sigma_{i=1}^{N} p(y_i | x_i \beta)$
↳ allows for "α" to select β

- can have multiple gaussian for creation of pdf