2 WEEK 1 SEMESTER 2

19/1/16. APLIED LINEAR STATISTICAL METHODS 2

Other name : Generalised linear models (GLM)

Introduction
Linear regression and how it can be generalised

Usually given a set of data:
$\{ y^{(i)}, x^{(i)} \}$ $i = 1, \dots, n$
$x^{(i)}$ ↑ vector called the explanatory variable
$y^{(i)}$ ↑ $y \in \mathbb{R}$ response variable

Data

for each variable defining $y^{(1)}$ $x^{(1)}$ → variable y_1 vector x_1
$y^{(2)}$ $x^{(2)}$ → y_2 x_2
$y^{(n)}$ $x^{(n)}$ → y_n x_n
Associate variable and vector

For each data point associate a variable / vector

LR: linear relationship $y_i = \beta^T x_i + \varepsilon_i$ (response, vector, error)
3 parameters
β^T - set of parameter
ε - noise in equation

$$y_2 = \beta^T x_2 + \varepsilon_2$$

$$y_n = \beta^T x_n + \varepsilon_n$$

Error $\varepsilon_i \sim N(0, \sigma^2)$ $\forall i = 1, \dots n$
$\varepsilon_i \perp \varepsilon_j$ $i \neq j$ (residuals) independent)

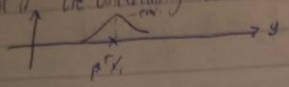
First equation $y_i = \beta^T x_i + \varepsilon_i \leftarrow N(0, \sigma^2)$
$(y_1^{(1)})$ only one observation available for y_1
$y_i^{(1)}$ only one obs available for it

Question: what is $p(y_i | \beta, x_i)$?

p: probability density function or distribution

$|\beta, x_i$: given β and x_i (know the value)

⇒ what is the uncertainty associated with y_i ?



Will be a translated normal dist with mean $\beta^T x_i$

$$\begin{aligned} \mathbb{E}[y_i] &= \mathbb{E}[x_i^T \beta + \varepsilon_i] \\ &= \mathbb{E}[x_i^T \beta] + \mathbb{E}[\varepsilon_i] \quad \text{taken as given} \\ &= x_i^T \beta + 0 \end{aligned}$$

$$Var[y_i] = \sigma^2$$

$$\begin{aligned} p(y_i | \beta, x_i) &= N(y_i, \beta^T x_i, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{(y_i - \beta^T x_i)^2}{2\sigma^2}\right] \end{aligned}$$

Can redefine Linear Regression as

response y_i is normally distributed such that: $p(y | x, \beta)$

$$\mathbb{E}[y_i] = \beta^T x_i \qquad \mathbb{E}[(y_i - \beta^T x_i)^2] = Var[y_i] = \sigma^2$$

GLM

$p(y | \beta x)$ is a member of one the exponential family of distributions.

- LR → One element is Normal/Gaussian distribution $y \in \mathbb{R}$
- Logistic Regression → Bernoulli distribution 0 or 1 outcome $y \in [0, 1]$
- → Binomial distribution generalisation of bernoulli $0 \to N$ integer number $y \in [1 \dots n]$
- Poisson Regression → Poisson Distribution discrete, can go to infinity $y \in I$

19/1/16 APPLIED LINEAR STATISTICAL METHODS II

↳ Exponential Distribution — models $y \in \mathbb{R}^+$ outcomes - positive real number like time to failure
Weibull Distribution ← survival analysis

Exponential Family of Distributions

$p(y|\theta) = \exp[a(y)\,b(\theta) + c(\theta) + d(y)]$

Function a of parameter y — Function c of θ

"Any distribution that you can write down in this form is a member of family"

Normal Distribution $(y|\theta = \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-y+\mu)^2/2\sigma^2$

$\exp\left[\frac{-(y-\mu)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right]$ — θ associated with application of y

$\exp\left[\underbrace{\frac{-y^2}{2\sigma^2}}_{d(y)} + \underbrace{\frac{2y\mu}{2\sigma^2}}_{a(y)\,b(\theta)} + \overbrace{\frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)}^{c}\right]$

$E[y]$ is defined w.r.t. θ.

Exponential Distribution

$p(y|\theta) = \theta \exp(-\theta y)$ — $y \in \mathbb{R}^+$ and not zero; $\theta \in \mathbb{R}^+$

$= \exp(-\theta y + \underbrace{\log\theta}_{c(\theta)})$

$b(\theta)\ a(y)$ — $d(y) = 0$

pdf must integrate to 1

$\int_0^{+\infty} p(y|\theta)\,dy = 1$

$\int_0^{+\infty} \theta \exp(-\theta y)\,dy = \left[-\exp^{-\theta y}\right]_0^{\infty}$

$= -\exp(-\theta\infty) + \exp(0)$

$0 + 1$

What is $\mathbb{E}[y]$? $= \int_0^{+\infty} y\, p(y|\theta)\, dy$ y time the function

$= \int_0^{\infty} y \cdot \theta \exp(-\theta y)\, dy$

$\mathbb{E}[y] = 1/\theta$ θ cannot be zero

Weibull Distribution $y \in \mathbb{R}^+$ $\theta \in \mathbb{R}^{+*}$ $\lambda \in \mathbb{R}^{+*}$

$p(y|\lambda\theta) = \lambda\theta y^{\lambda-1} \exp(\theta y^{\lambda})$

when $\lambda = 1$ weibull = exponential dist

Poisson Distribution ~~$y \in \mathbb{R}$~~ $y \in \mathbb{N}$ $\theta \in \mathbb{R}^{+*}$

$p(y|\theta) = \frac{\theta^y}{y!} \exp(-\theta)$

$\mathbb{E}[y] = \sum_{y \in \mathbb{N}} y \frac{\theta^y}{y!} \exp(-\theta) = \theta$

$= \exp(-\theta)\, \theta \sum_{y=1}^{\infty} \frac{\theta^{y-1}}{(y-1)!}$ Taylor expansion

$= \exp(-\theta) \exp(\theta)$

$= \theta$

θ is where information about explanatory variable will be insert

θ is the parameter used to plug in information from explanatory variable x

$\theta \rightarrow (x^T\beta)$ Somehow associated.

Example: In normal distribution $\mathbb{E}[y] = \mu = \theta = x^T\beta$ (linear regression)

$\theta = g(x^T\beta)$ (Link function, identity function in this case)

$\mathbb{R} \xrightarrow{\text{ATO}} \mathbb{R}$ mapping to different domain.

will want to map to outcome domain like exp and log, must be invertible

g and g^{-1} need to be overwritten

9/1/16 APPLIED LINEAR STATISTICAL METHODS 2

GLM

1. Have a set of independent responses y_i with associated explanatory variable x_i $\{(x_i, y_i)\}_{i=1,\dots,n}$

2. Response $y_i \sim p(y_i | \theta_i)$ member of the exponential family

3. Model construction $g(\mathbb{E}[y]) = x_i^T \beta$ (link function)
   $\mathbb{E}[y_i] = g^{-1}(x_i^T \beta)$

4. g: monotonic differentiable function (ensures inverse existence)

5. Estimate β by $\beta = \arg\max$ likelihood or argmax posterior probability

6. $\mathbb{E}[y] = g^{-1}(x^T \beta)$

21/9/16 APPLIED LINEAR STATISTICAL METHODS 2

y response variable $\dim(y) = 1$ 1 dimension

$\mathbb{E}[y] \Rightarrow g(x^T\beta)$ (related; link function)

$y \sim P_{y|\theta}(y|\theta)$

probability density function / distribution

continuous r.v discrete r.v

Example: In L.R using N dist

$y \sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(y - x^T\beta)^2}{2\sigma^2}\right)$ — g: identity function

In general: $y \sim P_{y|\theta} \rightarrow P_{y|x,\beta}$

some dist: weibull, binom, poisson.

constrained model

e.g. y grade from x hours put in with parameters β.

Data

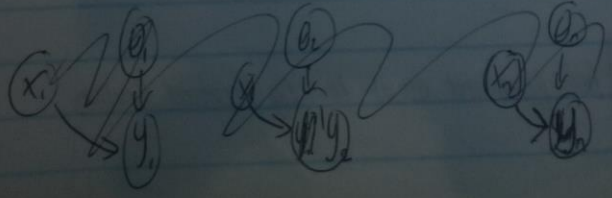
Say n observation hours, mark in exam → Variable (Definition of variables)

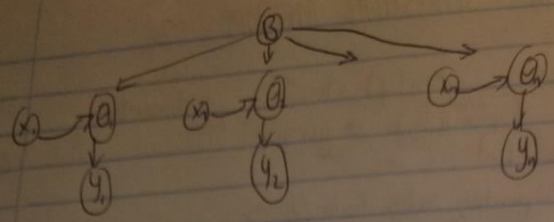
$x^{(1)}$	$y^{(1)}$	→	y_1 x_1 →	$y_1 \sim P_{y\|\theta}(y_1\|\theta)$ (Parameterised by θ)
$x^{(2)}$	$y^{(2)}$	→	y_2 x_2 →	$y_2 \sim P_{y\|\theta}(y_2\|\theta)$
$x^{(n)}$	$y^{(n)}$	→	y_n x_n →	$y_n \sim P_{y\|\theta}(y_n\|\theta)$

Value usually carried in note/book $\int P_{y|\theta}(y_1|\theta)\,dy_1 = 1$ (mean value)

$\int P_{y|\theta}(y^{(i)}|\theta)\,dy^{(i)} = 1$

Can do this!





Control θs by β
To estimate β wisely design a cost function

Likelihood ≡ Cost function corresponds to joint dist of all value given params needed
$P(y_1, y_2 \ldots y_n \mid \theta_1, \theta_2, \ldots \theta_n)$ i.e your model is independent.
$= \prod_{i=1}^{n} P(y_i \mid \theta_1, \theta_2, \theta_n)$ independence of response variables given the θs.

$= \prod_{i=1}^{n} P(y_i \mid \theta_i)$ only concerned with θ_i, other θ don't give information for y_i.

If: $y_i \sim P_{y|\theta}(y_i \mid \theta_i)$
Likelihood: $P(y_1, y_2 \ldots y_n \mid \theta_1, \ldots \theta_n) = \prod_{i=1}^{n} P_{y|\theta}(y_i \mid \theta_i)$ Joint density function

Take Log Likelihood: $\sum_{i=1}^{n} \log\left(P_{y|\theta}(y_i \mid \theta_i)\right)$
member of exponential family of distributions

If $P_{y|\theta}(y|\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$

Loglike: $\sum_{i=1}^{n} a(y_i)b(\theta_i) + c(\theta_i) + d(y_i)$

each $\theta_i \rightarrow g^{-1}(x_i^T\beta)$ link function

$\hat{\beta} = \arg\max$ loglike find param that maxes the Log likelihood function

2 21/a/16 APPLIED LINEAR STATISTICAL METHODS 2 3

GLMs : 1970s didn't have computational power

LR: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$= \underset{\text{vector}}{(1\ x_i)} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \varepsilon_i$

vector x_i β

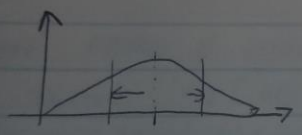
$y_i = \beta^T x + \varepsilon_i$

$\swarrow g$ used to get θ_i using link function

θ_i

can use to predict new value

$\hat{\beta} \longrightarrow p(y \mid x, \beta)$ $(y \mid x, \beta)$



can get CI from it.

Only have one maximum for cost function (convex function)
differentiate and equate to 0 for max value Only one global max solution

26/9/16 APPLIED LINEAR STATISTICAL METHODS 2.

LOGISTIC REGRESSION (Ch 7 in Dobson Book)

Usually: $\{(x_i, y_i)\}_{i=1,\dots,n}$.

For example: $\{(x_i, y_i, n_i)\}_{i=1,\dots,n}$ [Dose(x) #Exposed(n), #Killed(y)]

Binomial Distribution: $P(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

$y \in \{0,1,\dots,n\}$ $\theta \in [0,1]$

$\binom{n}{y} = \frac{n!}{(n-y)!\,y!}$ $\sum_{y=0}^{n}\binom{n}{y}\theta^y(1-\theta)^{n-y}$

$= (\theta+1-\theta)^n$ ← Pascal's triangle

$= 1^n = 1$

Binomial: member of exponential family of distributions?

↳ check $\exp[a(y)b(\theta) + c(\theta) + d(y)]$

if we can get binomial dist in this form it proves it is part of exponential family

$P(y|\theta) = \exp[\log[\binom{n}{y}\theta^y(1-\theta)^{n-y}]]$

$= \exp[\log\binom{n}{y} + y\log\theta + (n-y)\log(1-\theta)]$

$= \exp[\log\binom{n}{y} + y\log(\frac{\theta}{1-\theta}) + n\log(1-\theta)]$

$d(y)$ $a(y)b(\theta)$ $c(\theta)$

Value of θ that maximises $P(y|\theta)$?

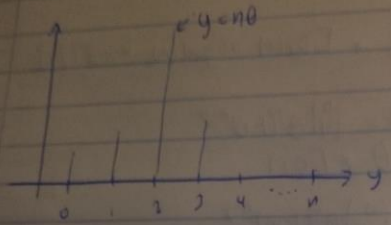
$\frac{\partial P(y|\theta)}{\partial\theta} = \binom{n}{y}\left[y\theta^{y-1}(1-\theta)^{n-y} - \theta^y(n-y)(1-\theta)^{n-y-1}\right]$

$= \binom{n}{y}\theta^{y-1}(1-\theta)^{n-y-1}\left[y(1-\theta) - \theta(n-y)\right]$

$\frac{\partial P(y|\theta)}{\partial \theta} = 0$ $\quad y(1-\theta) - \theta(n-y) = 0$

$y - y\theta - \theta n + \theta y = 0$

$y = n\theta$ or $\theta = y/n$



$E[y]$? When $y \sim P(y|\theta)$ (Binomial)

$$E[y] = \sum_{y=0}^{n} y \cdot P(y|\theta) = \sum_{y=0}^{n} y \cdot \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

$$= \sum_{y=1}^{n} y \cdot \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

because when $y=0$ the expression would have been 0.

$$y\binom{n}{y} = \frac{yn!}{(n-y)!\,y!} = \frac{y \cdot n \cdot (n-1)!}{(n-1-(y-1))!\,y\cdot(y-1)!} = n\binom{n-1}{y-1}$$

$$= n\theta \underbrace{\sum_{y=0}^{n-1} \binom{n-1}{y-1} \theta^{y-1} (1-\theta)^{(n-1-(y-1))}}_{\text{Bin}(\theta, n-1)} = n\theta$$

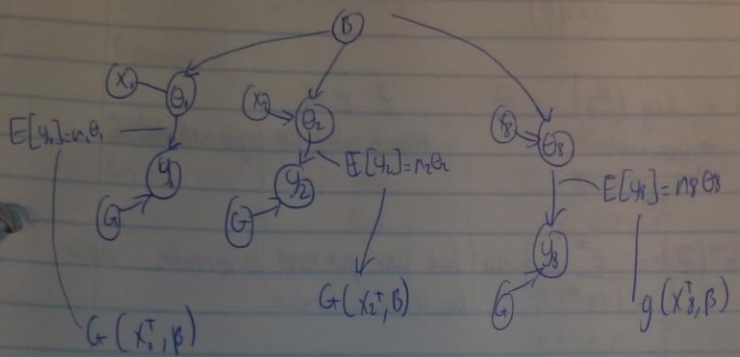
$y_1 \sim \text{Bin}(\theta_1, n_1)$

$y_2 \sim \text{Bin}(\theta_2, n_2)$

$y_3 \sim \text{Bin}(\theta_3, n_3)$

16/01/16 APPLIED LINEAR STATISTICAL METHODS 2

Saturated Model / Unrestricted GLM g-link function



$$L(\theta_1, \theta_2, \ldots, \theta_8) = P(y_1, y_2, \ldots y_8 \mid \theta_1, \theta_2 \ldots \theta_8)$$

$$= \prod_{i=1}^{8} P(y_i \mid \theta_i) \quad \text{(independence)}$$

$$= \prod_{i=1}^{8} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{(n_i - y_i)}$$

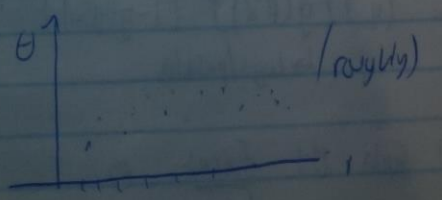
$$(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_8) = \arg\max L(\theta_1, \ldots, \theta_8)$$

$\hat{\theta}_1 = y_1/n_1$

$\hat{\theta}_2 = y_2/n_2$

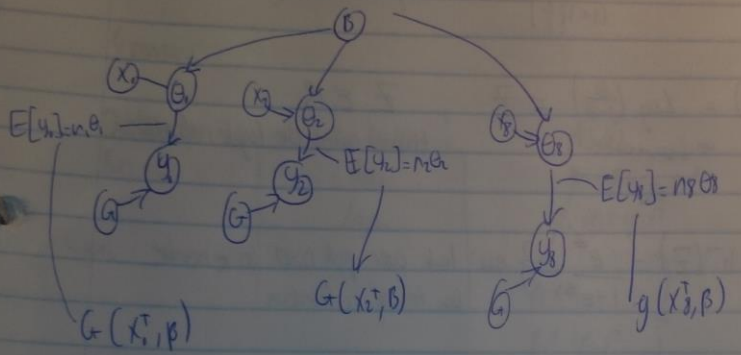
$\hat{\theta}_8 = y_8/n_8$

Can plot θ_i vs x_i



26/01/16 APPLIED LINEAR STATISTICAL METHODS 2

Saturated Model / Unrestricted GLM g-link function



$$\mathcal{L}(\theta_1,\theta_2,\ldots,\theta_8) = P(y_1,y_2,\ldots y_8 \mid \theta_1,\theta_2,\ldots\theta_8)$$

$$= \prod_{i=1}^{8} P(y_i \mid \theta_i) \quad \text{(independence)}$$

$$= \prod_{i=1}^{8} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{(n_i-y_i)}$$

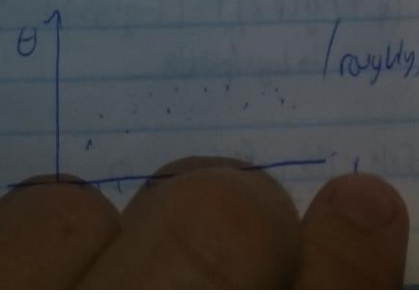
$$(\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_8) = \operatorname{argmax}\, \mathcal{L}(\theta_1,\ldots,\theta_8)$$

$\hat{\theta}_1 = y_1/n_1$

$\hat{\theta}_2 = y_2/n_2$

$\hat{\theta}_8 = y_8/n_8$

Can plot θ_i vs x_i



$\theta \in [0,1] \xrightarrow{g} X^T\beta \in \mathbb{R}$

$\xleftarrow{g^{-1}} \beta_0 + \beta X$ or $(1\times)\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix}$

Have to be invertible to translate back through the function to $[0,1]$

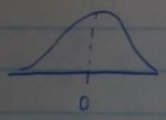
$b(\theta) = \text{Log}\left(\frac{\theta}{1-\theta}\right) = Z, \quad Z \in \mathbb{R}$

$[0,1] \Rightarrow (-\infty, +\infty) = \mathbb{R}$

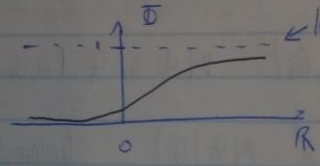
This is called the Logit function

b^{-1}? $b^{-1}(Z) = \frac{e^Z}{1+e^Z}$ } nice! look over put notes to remember how to invert a function

Normal Distribution



$\Rightarrow$



← Integral of normal distribution between 0 and 1

The above is called the probit and it is positive and strictly increasing, therefore the inverse exists i.e $\Phi \Rightarrow \Phi^{-1}$

$P(y|\beta x) = \binom{n}{y} (g(\beta^T x))^y (1-g(\beta^T x))^{n-y}$

$\mathcal{L}\beta = \prod_{i=1}^{n} P(y_i|\beta x_i)$

$= \prod_{i=1}^{n} \binom{n_i}{y_i} [g(\beta^T x_i)]^{y_i} (1-g(\beta^T x_i))^{n_i - y_i}$

↳ Logit/probit/etc

Find $\hat{\beta}$ such that $\frac{\delta \mathcal{L}(\beta)}{\delta(\beta)} = 0$

26/01/16 APPLIED LINEAR STATISTICAL METHODS 2

Note: You could add a $\beta_2 X^2$ if you wanted to add another dimension to your model

Summary:

Distribution	Link	Model
Binomial	Probit	$\beta_0 + \beta_1 X$
	Logit	$\beta_0 + \beta_1 X + \beta_2 X^2$
		β_0
		$\beta_1 X$
		$\beta_1 X + \beta_2 X^2$
		$\beta_0 + \beta_2 X^2$

} in model creation

- Can use AIC & BIC to determine the "best" model
- Also may calculate the deviance to find out if the model is any good in the first place

2/8/16 APPLIED LINEAR STATISTICAL METHODS 2

GLM Code Study

y_i: did patient survive? 1: NO 0: yes

$\{(x_i, y_i)\}$ $i = 1, \ldots, 40$ $n = 40$

x_i: age $y_i \in \{0, 1\}$

Response Variable $y_i \sim$ Binomial distribution

$P(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$

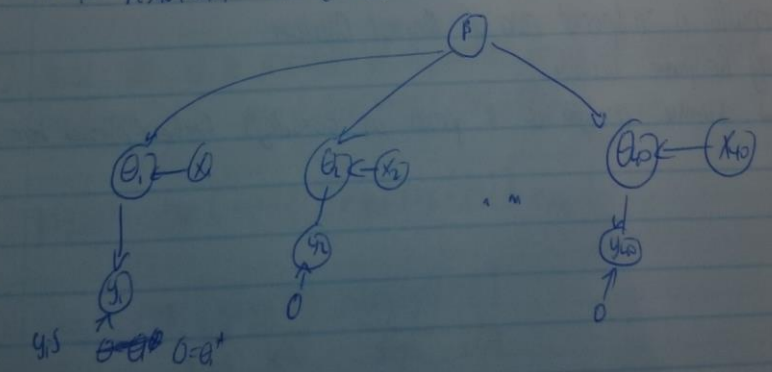
$i = 1$ $P(y_1|\theta) = \binom{1}{y_1} \theta_1^{y_1}(1-\theta_1)^{1-y_1}$ $y_i \in \{0,1\}$

$= \theta_1^{y_1}(1-\theta_1)^{1-y_1}$

[When there is only 1 choice (binary) the distribution used is called the bernoulli distribution]

$i = 2$ $P(y_2|\theta_2) = \theta_2^{y_2}(1-\theta_2)^{1-y_2}$

$i = 40$ $P(y_{40}|\theta_{40}) = \theta_{40}^{y_{40}}(1-\theta_{40})^{1-y_{40}}$



y_is ~~$\theta = \theta^*$~~ $\theta = \theta^*$

0 ⇒ likelihood solution of selected/current model with $n = 1$ (bernoulli)

$E[y] = n\theta = \theta$ $n = 1$ (bernoulli)

$\frac{\delta P(y|\theta)}{\delta(\theta)} = 0 \Rightarrow \hat{\theta} = y/n = y$

$L(\theta_1, \theta_2 \ldots \theta_m) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}$

$\dim(\theta^*) = 4$

$(\theta_1^* \; \theta_2^* \; \theta_3^* \; \theta_4^*) = \operatorname{argmax} L(\theta_1, \theta_2, \ldots, \theta_m)$

Link fn: $b(\theta) = g(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ [Logit]

From $p(y|\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$

GLM:

$L(\beta) = \prod_{i=1}^{n} g^{-1}(\beta_0 + \beta_1 x_i)^{y_i}[1 - g^{-1}(\beta_0 + \beta_1 x_i)]^{1-y_i}$

$\dim(\beta) = 2 \Rightarrow$ collaborating the model

To test if a relationship exists in the model:

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$

Summary:

- Bernulli is a special case of Binomial Distribution
- Binary Response Variable
- Follow similar analysis of & proofs as tuesday's lecture (previous notes)

2

1

02/02/16 APPLIED LINEAR STATISTICAL METHODS 2

Poisson Distribution - Poisson Regression

$$P(y|\lambda) = \frac{\lambda^y}{y!}\exp(-\lambda) \qquad y \in N \qquad \lambda > 0$$

Exercise:

1. Show this a distribution How? Show it is always positive

$P(y|\lambda) > 0 \Rightarrow$ positive function

2. Show $\sum_{y=0}^{\infty} p(y|\lambda) = 1$

$$= \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}\exp(-\lambda) \qquad = \exp(-\lambda)\sum_{y=0}^{\infty}\frac{\lambda^y}{y!}$$

we know it will integrate to 1

A side: Taylor expansion $\exp(x) = \sum_{n=0}^{\infty}\frac{x^n}{n!}$ $\Rightarrow$ this is what we have above

Taylor $\exp(\lambda) = \sum_{y=0}^{\infty}\frac{\lambda^y}{y!}$

2. Show this is a member of exponential family of distributions.
In form of: $\exp[a(y)b(\lambda) + c(\lambda) + d(y)]$

$$p(y|\lambda) = \exp[\text{Log}(\lambda^y)]\exp[\text{Log}(\tfrac{1}{y!})]\exp(-\lambda)$$

$$= \exp[\underbrace{y}_{a(y)}\underbrace{\text{Log}(\lambda)}_{b(\lambda)} \underbrace{-\text{Log}(y!)}_{d(y)} \underbrace{-\lambda}_{c(\lambda)}]$$

3. Compute the expectation of y and verify this is a function of λ.

$$\mathbb{E}[y] = \int y\, p(y|\lambda)\, dy$$ ↓ change to sum because y is a discrete r.v.

$$= \sum_{y=0}^{\infty} y\, p(y|\lambda)$$

$= \sum_{y=0}^{\infty} \frac{y}{y!} \lambda^y \exp(-\lambda)$ take out $\exp(-\lambda)$ as doesn't depend on y

$= \exp(-\lambda) \sum_{y=0}^{\infty} \frac{y \lambda^y}{y!}$

$= \exp(-\lambda) \sum_{y=0}^{\infty} \frac{y \lambda^y}{y(y-1)!}$ Start Sum of $y=1$ as at $y=0$ Sum $=0$

$= \exp(-\lambda)(\lambda) \boxed{\sum_{y=1}^{\infty} \frac{y \lambda^{y-1}}{y(y-1)!}}$

$\sum_{y=1}^{\infty} \frac{\lambda^{y-1}}{(y-1)!} = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \exp(\lambda)$ (taylor expansion again)

$= \exp(-\lambda)(\lambda)\exp(\lambda)$

$= \lambda$ = expectation

4. Compute the maximum w.r.t λ of the distribution

Find λ such that $\frac{dl}{d\lambda} = 0$

$\frac{dl}{d\lambda} = \frac{1}{y!}\left[y\lambda^{y-1} \exp(-\lambda) - \exp(-\lambda)\lambda^y \right]$ (u, v)

$= \frac{\lambda^{y-1}\exp(-\lambda)}{y!}\left[y - \lambda \right]$

equal to 0 when $\lambda = y$ (from second bracket)

Solved solve of Max likihood is $\lambda = y$.

y is an integer, λ is in $\mathbb{R}^+$

Eg map $\mathbb{R}$ onto $\mathbb{R}^+$, can use it as a link between our function and expectation

$E[y] = \lambda \in \mathbb{R}$ $\quad \log(\lambda) = x^T\beta$ $\quad x^T\beta$

2.

only

02/02/16 APPLIED LINEAR STATISTICAL METHODS 2

3

$E[y] = X \in \mathbb{R}$ $\xrightarrow{Log(\lambda) = x^T\beta}$ $x^T\beta$ $\xleftarrow{\lambda = \exp(x^T\beta)}$ Link function

Are some links with poisson and binomial
Use poisson for number of occurances of an event
Sometimes have n poisson trials.

Binomial $P[y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$
usually given observation (y_i, x_i, n_i) $E[y|\theta] = n\theta$

If we use Log as link function $\log[E_{poi}[y]] = \log(\lambda) = x^T\beta$
→ two possible models for r.v y (poisson or binomial).

Instead ⇒ $\log\left[E[y]/n\right] = \log\left(\frac{\lambda}{n}\right) = x^T\beta$ like normalising (using size of group)
$\log(\lambda) = \log(n) + x^T\beta$ Divide expectation by size n.
equivalent of writing ↑ offset

Look at R code example
Same strategy as last week.

n - is called exposure, "how many beetle are you exposing to drug?"
log(n) is the offset. ⇒ only affect β_0 (intercept) in model

5pm 1.20 Arnold Duey a Stats clinic

4/02/16

Beetle Distribution	Data	Link	Explanatory Model	AIC (ps)
Binomial	Probit AIC: 40.32 Logit AIC: 41.91	$\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$	$\beta_0+\beta_1\vec{x}$; $\beta_0+\beta_1\vec{x}+\beta_2 x^2$	$-2\text{Log}\mathcal{L}(\hat{\beta}) + 2m$ complexity varied with # of β's
Poisson	Log AIC: 57.72			

Insurance Detailer Example: See R code on website

Show that:

$$\lim_{n\to\infty} \text{Bin}(y;\theta,n) = \text{Pois}(y;\lambda)$$

$$\mathbb{E}_{\text{Bin}}[y] = n\theta \qquad \mathbb{E}_{\text{Pois}}[y] = \lambda$$

$$\lim_{n\to\infty} \binom{n}{y}\theta^y(1-\theta)^{n-y} \stackrel{?}{=} \frac{\lambda^y}{y!}\exp(-\lambda) \quad (?)$$

$\theta = \lambda/n$

$$\binom{n}{y}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} \xrightarrow{\text{expand}} \frac{n!}{n^y(n-y)!}\cdot\frac{\lambda^y}{y!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-y}$$

$A(n)$ → $\to 1$; $B(n) \to \exp(-\lambda)$; when $n\to\infty$, y does not

$\lim_{n\to\infty} A(n)$? $y = 1$ $\quad A(n) = \frac{n(n-1)\cdots(n-y+1)}{n\cdot n\cdots n} \to 1$ (doesn't depend on n; $= 1\cdot(1-\frac{1}{n})\cdots$)

$\lim_{n\to\infty} B(n)$? $\left(1-\frac{\lambda}{n}\right)^n = \exp(-\lambda) \Rightarrow \left(1-\frac{\lambda}{n}\right)^n = \sum_{k=0}^{n}\binom{n}{k}\left(\frac{-\lambda}{n}\right)^k$

$$= \frac{n!}{(n-k)!k!}\cdot\frac{-\lambda^k}{n^k}$$

$$= \frac{n!}{(n-k)!n^k}\cdot\frac{-\lambda^k}{k!}$$

Taylor expansion $\exp(-\lambda)$

$A(n) \to 1$ before

Put all back together

$$(1)\left(\frac{\lambda^y}{y!}\right)(\exp(-\lambda))(1) = \text{poisson distribution}$$

When 1 cdb $\Rightarrow$ Bernoulli

When $n\to\infty$ cdb $\Rightarrow$ Poisson Distribution

02/02/16 APPLIED LINEAR STATISTICAL METHODS 2

Looking at posterior odds

Example, oysters and dirty water, how many are infected?

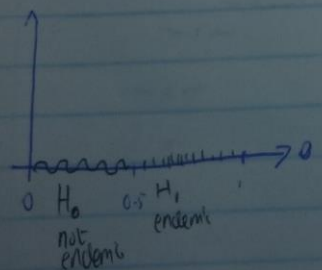
$n = 10$ people tested
$y = 7$ infected

$H_0: \theta \leq 0.5$ not endemic $H_1: \theta > 0.5$ Infection is endemic

Binomial:
$P(y|\theta) = \binom{10}{7} \theta^7 (1-\theta)^3$
$\theta_b^* = y/n = 0.7$ for Maximum Likelihood

How to make 95% CI for this? 0.7 value

Posterior dist of θ given y:
$P(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$, $p(y) = \int p(y|\theta) p(\theta)\, d\theta$ ← "normalising constant"

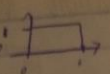


$P = \int_{0.5}^{1} p(\theta|y)\, d\theta \geq 0.95$ then H_1 is true

Distribution	$E[y]$	θ	Nuisance Parameter	Max likelihood θ^*	$a(y)\ b(\theta)\ c(\theta)\ d(y)$
Normal $\frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(y-\theta_g)^2}{2\sigma^2}\right]$	θ_g	θ_g	σ	$\theta_g^* = y$	Solution not unique for normal dist.
Binomial $\binom{n}{y} \theta_b^y (1-\theta_b)^{n-y}$	$n\theta_b$	θ_b	n	$\theta_b^* = \frac{y}{n}$	
Poisson $\frac{\theta_p^y}{y!} \exp[-\theta_p]$	θ_p	θ_p	none	$\theta_p^* = y$	

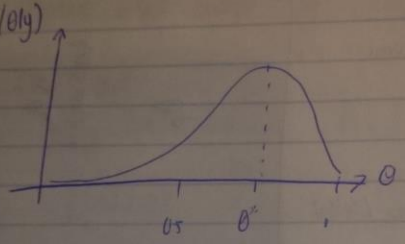
Bernoulli when $n=1$

1/02/16

Belief (A):  θ can be anywhere between 0 and 1 "uniform probability for value of θ"

have to integrate to 1 on y axis

θ MP = to solve: $\frac{\int_{0.5}^{1} \binom{10}{7} \theta^7 (1-\theta)^3 \, d\theta}{\int_{0}^{1} \binom{10}{7} \theta^7 (1-\theta)^3 \, d\theta}$ $P(\theta) = 1$



checking that integration from 0.5 → 1 is ≥ 95% to prove hypothesis

$\theta^7(1-\theta)^3 \rightarrow \frac{-\theta^{11}}{11} + \frac{3\theta^{10}}{10} - \frac{3\theta^9}{9} + \frac{\theta^8}{8}$ $\left(\frac{-3\theta^9}{9}\right)$

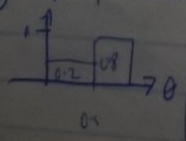
$(1-\theta)^3 = 1 - 3\theta + 3\theta^2 - \theta^3$

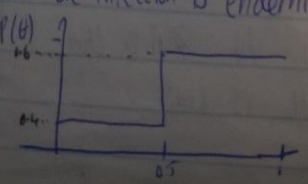
$\theta^7(1-\theta)^3 = \theta^7 - 3\theta^8 + 3\theta^9 - \theta^{10}$ integrated ⇒

$= \theta^8/8 - 3/9\,\theta^9 + 1/10\,\theta^{10} - \theta^{11}/11$

$\frac{[\ldots]_{0.5}^{1}}{[\ldots]_{0}^{1}} = 0.88 = \theta^*$

(Fun case): Belief (B): expert opinion: 80% sure that the infection is endemic





Compute the prior again (the integral fraction)

$$P(\text{expert}) = \frac{\int_{0.5}^{1} \binom{10}{7} \theta^7 (1-\theta)^3 \cdot 1.6}{\int_{0}^{0.5} \binom{10}{7} \theta^7 (1-\theta)^3 (0.4) + \int_{0.5}^{1} \binom{10}{7} \theta^7 (1-\theta)^3 (1.6)}$$

← split because no longer uniform

Check if this is 0.95 to confirm H_1

$$\theta^7 (1-\theta)^3 \, d\theta = \frac{\theta^{11}}{11} + \frac{3\theta^{10}}{10} - \frac{3\theta^9}{9} + \frac{\theta^8}{8}$$

Results in 0.9691

Expert saved money by not needing extra tests or larger n size to gain significant results

11/02/16 APPLIED LINEAR STATISTIAL METHODS

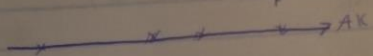
Choosing GLM Models

1. AIC For model selection

AIC = $-2\log\hat{L} + 2M$ Pick ↑smallest AIC

AIC Saturated $-2\log L_\theta + 2\dim(\theta)$

GLM $-2\log L_\beta + 2\dim(\beta)$



2. Deviance: good Model? ← Binomial Distribution (Saturated Model)

Deviance = $-2\log\left[\frac{L_\beta}{L_\theta}\right]$ ← Binomial constrained by $\beta^T x$

For example, binomial distribution with $\beta^T x$ and Logit link

Use $qchisq(0.95, 6)$ command in R for Chi-Square test on Deviance

$\propto$ DF $\dim(\theta - \beta)$

Relation Between AIC and Deviance

Example. $\text{Deviance}^{\text{Binomial}} = -2\log_\theta^{\text{Binomial}} + 2\log L_\beta^{\text{Binomial}}$

$= AIC_\theta - 2\dim\theta - AIC_\beta + 2\dim(\beta)$

$= AIC_\theta - AIC_\beta + 2(\dim(\beta) - \dim(\theta))$

$AIC_\theta = -2\log L_\theta^{Bin} + 2\,Dim(\theta)$

$AIC_\beta = -2\log L_\beta^{Binom} + 2\,Dim(\beta)$

Chi-Square Distribution

$Z \sim N(0,1)$ $i = 1, \ldots k$ independent i.i.d.

$X = \sum_{i=1}^{k} Z_i^2$ $X \sim \chi^2(k)$ prove this?

When $X = Z^2$

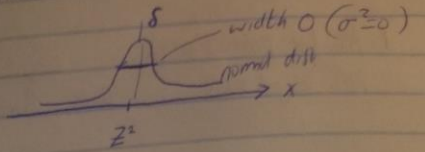
$P_X(x) = \int P_{XZ}(x, z)\,dz$ conditional distribution

02/16

No uncertainty in x when given z (x is z^2 remember)

Dirac distribution has variance 0: no uncertainty

$p(x|z) = \delta(x - z^2)$



$\int_{-\infty}^{\infty} \delta(x-z^2)\,dx = 1$ property of dirac function

$= \int \delta(x-z)^2 \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz$

Use Dirac property to rewrite dirac part

$\Rightarrow \frac{1}{2|\sqrt{x}|}\left[\delta(\sqrt{x}+z) + \delta(\sqrt{x}-z)\right]$

$\rightarrow \int \frac{1}{2|\sqrt{x}|}\left[\delta(\sqrt{x}+z) + \delta(\sqrt{x}-z)\right] \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right] dz$

$= \frac{1}{2|\sqrt{x}|}\int \delta(\sqrt{x}-z)\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz + \int \delta(\sqrt{x}+z)\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz$

(two) (two)

Next property: $\int_{-\infty}^{\infty} f(t)\,\delta(t-T)\,dt = f(T)$

$= \frac{1}{2|\sqrt{x}|}\left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x}{2}\right) + \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x}{2}\right)\right]$

Chi-Square: for degree k $\quad P_k(x) = \frac{1}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} \exp\left[\frac{-x}{2}\right] \quad x \geq 0$

Degree of freedom

$\mathbb{E}[x] = k$

Gamma

Look for Devore near value of k as an indicator

writing

Devore for binomial (k) poisson (k) etc

exam??

Re-check definition of chi-square dist may not be correct here

6/02/16 APPLIED LINEAR STATISTICAL METHODS

No lecture Thursday 8th feb

Smoking Dataset Example

Groups	Response Variable	"nuisance parameter" (in general can be size of group)	x^{age}	$x^{smoking}$	$\hat{\theta}$ Sample Proportion
$i=1$	$y_1 = 32$	$n_1 = 52407 = 1$	35-44	1= Yes	$\hat{\theta}_1 = \frac{32}{52407}$
$i=2$	$y_2 = 104$	$n_2 = 43248 = 2$	45-54	1= Yes	
$i=3$	$y_3 = 206$	$n_3 = 28612 = 3$	55-64	1= Yes	$\hat{\theta}_3 = \frac{206}{28612}$
$i=4$	$y_4 = 186$	$n_4 = 12663 = 4$	65-74	1=Yes	
$i=5$	$y_5 = 106$	$n_5 = 5317 = 5$	75-84	1=Yes	
$i=6$	$y_6 = 2$	$n_6 = 18790 = 1$	35-44	0=No	
$i=7$	$y_7 = 12$	$n_7 = 10673 = 2$	45-54	0=No	
$i=8$	$y_8 = 28$	$n_8 = 5710 = 3$	55-64	0=No	
$i=9$	$y_9 = 28$	$n_9 = 2583 = 4$	65-74	0=No	
$i=10$	$y_{10} = 31$	$n_{10} = 1462 = 5$	75-84	0=No	
		or take middle value of category	↑ Ordinal	↑ nominal	

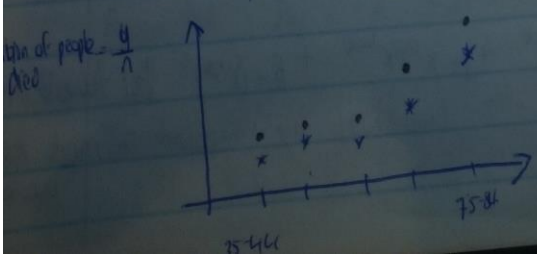
$N = 10$ (10 response variables)

Nature of explanatory variable:

* nominal (eg. red, blue, green) [e.g dead, alive or male, female] no order
* ordinal Natural order between states of x.
* continuous e.g weight, age, temperature

For the $\hat{\theta}$'s, draw a graph of all value to see difference

Separate into 2 groups



16/02/16 APPLIED LINEAR STATISTICAL METHODS

No lecture Thursday 8th feb

Smoking Dataset Example

Groups	Response Variable	"number present" (in general can be size of group)	X^{age}	$X^{smoking}$	$\hat{p}_{smokers}$
$i=1$	$y_1 = 32$	$n_1 = 52407 = 1$	35-44	1= Yes	$\hat{p}_1 = \frac{32}{52407}$
$i=2$	$y_2 = 104$	$n_2 = 43248 = 2$	45-54	1= Yes	
$i=3$	$y_3 = 206$	$n_3 = 28612 = 3$	55-64	1= Yes	$\hat{p}_3 = \frac{206}{28612}$
$i=4$	$y_4 = 186$	$n_4 = 12663 = 4$	65-74	1=Yes	
$i=5$	$y_5 = 102$	$n_5 = 5207 = 5$	75-84	1=Yes	
$i=6$	$y_6 = 2$	$n_6 = 18790 = 1$	35-44	0=No	
$i=7$	$y_7 = 12$	$n_7 = 10673 = 2$	45-54	0=No	
$i=8$	$y_8 = 28$	$n_8 = 5710 = 3$	55-64	0=No	
$i=9$	$y_9 = 28$	$n_9 = 2585 = 4$	65-74	0=No	
$i=10$	$y_{10} = 31$	$n_{10} = 1462 = 5$	75-84	0=No	
		or some middle value of category	ordinal ↑	nominal ↑	

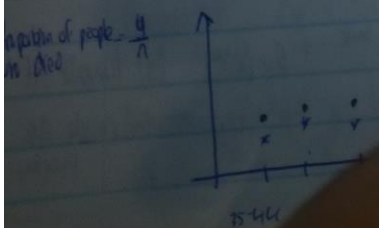
$N = 10$ (10 response variables)

Nature of explanatory variable:

* nominal (eg. red, blue, green) [e.g dead, alive or male, female] no order
* ordinal natural order between states of x
* continuous e.g weight, age, temperature

For the $\hat{\theta}$'s, draw a graph of all value to see difference

Seperate into 2 group



02/16

GLM tries to fit a curve between these points

Need to determine $\beta^T x$

Could take age^2 as an explanatory variable ⇒ 1, 2, 3, 4, 5
⇒ 1, 4, 9, 16, 25

Smoking is a binary variable. $smoke^2$ has no effect

Multiply smoke by age (Product) ⇒ 1, 2, 3, 4, 5, 0, 0, 0, 0, 0 → mixed effects model

Multiply age^2 x Smoke ⇒ 1, 4, 9, 16, 25, 0, 0, 0, 0, 0.

when the variables add

Point ⇒ given two explanatory variables, can make new variables if relationship is not linear.

normal. Binary

$\beta_0 + \beta_1 x^{age} + \beta_2 x^{smoke} + \beta_3 x^{age} \cdot x^{smoke} + \beta_4 (x^{age})^2 + \beta_5 (x^{age})^2 x^{smoke}$

Largest model from what we have defined in the table

When $x^{smoke} = 1$: $(\beta_0 + \beta_2) + (\beta_1 + \beta_3) x^{age} + (\beta_4 + \beta_5)(x^{age})^2$ "mixed model"

$x^{smoke} = 0$: $\beta_0 + \beta_1 x^{age} + \beta_4 (x^{age})^2$

6 parameters here

Binomial OR poisson is the choice of distribution

Try all possible models and choose model with lowest AIC

Dist	Link	$X^T\beta$
Binomial	logit	
	Probit	
Poisson	Log	

Fitted model: $\beta_0 + \beta_1 x^{age} + \beta_2 x^{smoke} + \beta_3 x^{age} x^{smoke} + \beta_4 (x^{age})^2$ had lowest AIC

AIC: Poisson + Log link function 667

Binomial + logit 666.3

Binomial + probit = 666.3

Tried with Age^2 removed ⇒ made AIC worse

16/02/16 APPLIED LINEAR STATISTICAL METHODS

When mixture effect term was removed, AIC increased

AIC: Poisson+Log : 667
Binomial+Logit : 668

Write the GLM $g(\theta) = \beta^T x$

Poisson+Log:

$$Log(\theta) = \hat{\beta}_0 + \hat{\beta}_1 X^{age} + \hat{\beta}_2 X^{Smoke} + \hat{\beta}_3 X^{age} X^{Smoke} + \hat{\beta}_4 (X^{age})^2$$

Binomial logit $Log\left(\frac{\theta}{1-\theta}\right) = \hat{\beta}_0 + \hat{\beta}_i \ldots\ldots$ $\hat{\beta}$'s will be different because of different Likelihood calculation

Deviances? $-2Log\left[\frac{Saturated\ (L)}{\lambda\ GLM}\right]$

Term $-2Log(L)$

1. For Saturated model - poisson dist
   - Binomial
2. For GLM Model Poisson+Log
3. For GLM Binomial+Logit.

$$Poisson(y|\lambda) = \frac{\lambda^2}{y!} exp(-\lambda)$$ (10 groups N=10)

$$L(\lambda_1, \lambda_2 \ldots \lambda_{10}) = \prod_{i=1}^{10} \frac{\lambda_i^{y_i}}{y_i!} exp(-\lambda_i)$$

λ is expectation not proportion
y_i = dead people in each group i.

$$Log(L) \Rightarrow \sum_{i=1}^{10} y_i Log\,\lambda_i - \lambda_i - Log(y_i!)$$

$\hat{\lambda}_i = y_i$ Saturated model (Max likelihood Solution)

Deviance/AIC $Log\,L(\hat{\lambda}_1 \ldots \hat{\lambda}_{10})$

$$Binomial\ p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

$$L(\theta_1 \ldots \theta_{10}) = \prod_{i=1}^{10} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta)^{n_i - y_i}$$ (applied to each group)

02/16

ML Solution $\hat{\theta} = y/n$ (Saturated)
Deviance/AIC: $-2\log(L)(\hat{\theta}_i, \theta_{i0})$

Binomial
$L(\theta, \theta_{i0}) = \prod_{i=1}^{10} \binom{n}{y} \theta_i^{y_i} (1-\theta_i)^{n-y}$
Saturated replace θ with $\hat{\theta}_i = y_i/n_i$

Poisson
$L(\lambda, \lambda_0) = \prod_{i=1}^{10} \frac{\lambda^{y_i}}{y_i!} \exp[-\lambda_i]$
Saturated $\hat{\lambda}_i = y_i$
GLM (log) $\tilde{\theta}_i = \exp(\hat{\beta}^T x_i)$ likelihood only dependent on (β) $L(\beta)$
(saturated) $\frac{\tilde{\lambda}_i}{n_i} = \exp(\hat{\beta}^T x_i)$

Different related to age? look at the coefficients β's to determine
look at standard errors associated with the β's.
"could I remove age explanatory variable from model ⇒ ~~factor~~ variable is important"
Had to be used in the model

Look at R code online

23/02/16 APPLIED LINEAR STATISTICAL METHODS

Response Variable	Member of Exponential family
$y \in \{0,1\}$	Bernoulli
$y \in \{0,1,\dots,n\}$	Binomial
$y \in \mathbb{N}$ (integer)	Poisson
$y \in \mathbb{R}$	Gaussian
$y \in \mathbb{R}^+$	exponential, weibull

positive real numbers (like time to failure)

$E[y] \; \theta : g(x^T\beta)$

Leukemia example
Response variable y is time
Is placebo time same as drug time?
Response variable for duration and any positive real number

Survival Analysis
Modelling time until failure
Survival analysis is concerned with the statistical modelling of time to failure from a well defined origin (of time) or starting point
- time of hard drive to fail from the time it has been built/bought
- time of patient to die from the time disease has been diagnosed

1 Distribution.
- Times y are non-negative $(y \in \mathbb{R}^+)$
- Distributions are skewed with long tails
- Some subjects may survive beyond the study and their failure time is not observed. In this case, the data is said to be "censured."

Exponential Distribution

$p(y|\theta) = \theta \exp[-\theta y]$ $\quad y \in \mathbb{R}^+$ $\quad \theta \in \mathbb{R}^{+*}$ (excludes 0)

$\mathbb{R}^+ / \{0\}$

3/02/16

Expectation $E[y]$?

$E[y] = \int_{y \in \mathbb{R}} y\, p(y|\theta)\, dy$

$= \int_0^\infty y\, \theta \exp[-\theta y]\, dy$

$= \frac{1}{\theta}$ by integration by parts

positive number

Best link function to map θ to positive real number, $\mathbb{R}^+$ onto $\mathbb{R}$

g should be exponential g^{-1} should be Log

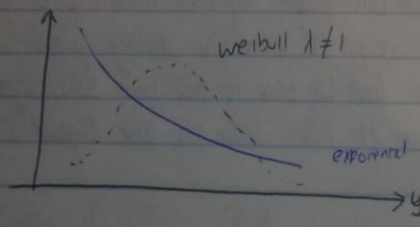
Weibull Distribution

$p(y|\theta,\lambda) = \lambda\theta y^{\lambda-1} \exp[-\theta y^\lambda]$ ← y power of λ

λ nuisance parameter

~~$y \in \mathbb{R}$~~ $y \in \mathbb{R}^+$ $\theta \in \mathbb{R}^{+*}$ $\lambda \in \mathbb{R}^{+*}$

Exponential Dist is a particular case of Weibull with $\lambda = 1$



$E[y]$? $E[y] = \int_0^\infty y\, \lambda\theta y^{\lambda-1} \exp[-\theta y^\lambda]\, dy$

$= \int_0^\infty \lambda\theta y^\lambda \exp[-\theta y^\lambda]\, dy$ *EXAM PROVE THIS

Show $E[y] = (\frac{1}{\theta})^{\frac{1}{\lambda}} \Gamma(1+\frac{1}{\lambda})$ gamma

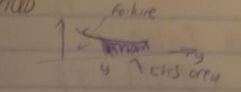
$\Gamma(\mu) = \int_0^\infty s^{u-1} \exp[-s]\, ds$ gamma

Survivor and Hazard function

(A) Probability of failure $P(0 \leq t \leq y) = \int_0^y p(t|\theta)\, dt$

p can be exponential or weibull area between 0 and y

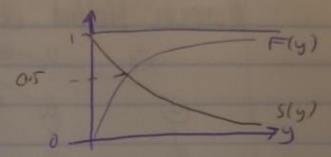
23/02/16 APPLIED LINEAR STATISTICAL METHOD



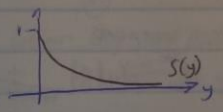
Probability of Survival beyond time y
$S(y) = 1 - F(y)$ (area to right of y)

B. Exercise $F(y)$, $S(y)$ for exponential distribution?

$F(y) = \int_0^y \theta \exp(-\theta t)\, dt$
$= [-\exp(-\theta t)]_0^y$
$= 1 - \exp(-\theta y)$ = probability of failure.



$S(y) = 1 - 1 + \exp(-\theta y) = \exp(-\theta y)$



At 0.5 some probability of failure or since
Later $F(y) = 0.5$ or $S(y) = 0.5$ called the median point of survival
$F(y) = S(y) = 0.5$

C. Hazard function

$h(y) = \frac{P(y \le t \le y+dy)}{P(y \le t)}$ chance of failure between time y and $y+dy$ / Surviving beyond time y $S(y)$

$= P(y|\theta)dy$

"Chance of dieing in next minute given a survival rate"

$h(y) = \frac{P(y|\theta)}{S(y)}$ pdf weibull or exponential function — deductive function

Cumulative hazard function $H(y) = -\log[S(y)]$ primitive

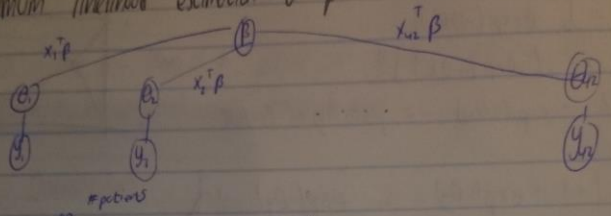
Link function

$g(E[y_i]) = x_i^T \beta$

↑

Log θ $\exp(x_i^T \beta)$

Maximum likelihood estimation of β.



$L(\beta) = \prod_{i=1}^{42} p(y_i|\theta)$ ← $\exp(x_i^T\beta)$ Weibull or exponential

This is the definition of the likelihood when y_i is uncensored.

When y_i is censored, use $S(y_i)$ instead of pdf $p(y_i|\theta_i)$

δ_i an indicator variable 0 if y_i is censored 1 if it is not $\delta_i \in \{0$ censored, 1 otherwise$\}$

↳ $L(\beta) = \prod_{i=1}^{42} p(y_i|\theta_i)^{\delta_i} S(y_i)^{1-\delta_i}$

↑ take either term depending on censored or not

- if you lose track of patient or survives duration of study

Example exponential Distribution. $L(\theta_1, \theta_{42}) = \prod_{i=1}^{42} [\theta_i \exp(-\theta_i y_i)]^{\delta_i} [\exp(-\theta_i y_i)]^{\delta_i}$

(Saturated model)

$L(\beta) = \prod_{i=1}^{42} [\exp(\beta^T x_i)]$

censoring variable given in example data

GLM not defined for exponential or weibull

use "Survival" package

23/02/16 APPLIED LINEAR STATISTICAL METHODS

DF = both = 2 in example of lukkemia
Scale is $1/\lambda$ value here

Is the drug better than the placebo effect?
look at
Median survival time $y_{med} = \frac{\log(2)}{\theta}$ for exponential dist

$= \frac{\log(2)}{\exp(\hat{\beta}^T x)}$ $(\beta_0 + \beta_1 x)$
↳ x = 0 if placebo
x = 1 drug

$\frac{y_{med}^{x=1}}{y_{med}^{x=0}} = \frac{1}{\exp\hat{\beta}_1}$ look at ratio between drug and placebo.
$= \exp(-1.53) = \frac{1}{4} \Rightarrow \frac{y_{med}^{x=0}}{4} = y_{med}^{x=1}$

$\frac{h^{x=0}}{4} = h^{x=1}$

4 times higher than treatment group. Probability is 4 times higher with placebo rate

4 times more likely to fail.

R output gives β's estimates on corresponding standard errors
± 2 SE
$2.32 \pm 2(0.4)$ 95% CI for $\beta_1 \Rightarrow \beta_1$ is not 0.

Same as in linear regression
relationship between time and drug exists

25/02/16 APPLIED LINEAR STATISTICAL

Exponential Distribution

$E[y] = \frac{1}{\theta}$

$\log[E[y]] = \log\left(\frac{1}{\theta}\right) = \beta_0 + \beta_1 x$ $\quad x \in \{0 \text{ placebo/control}, 1 \text{ drug}\}$

map $\mathbb{R}^+$ onto $\mathbb{R}$

$\log(\theta) = -\beta_0 - \beta_1 x$

$\theta = \exp[-\beta_0 - \beta_1 x]$

look at hazard function or median to see if drug is working

Weibull if is true for weibull → true for exponential

1. Show it is a distribution

y positive, θ positive, exp is positive → positive function

must show that it integrated to 1

$$\int_0^\infty \lambda\theta y^{\lambda-1} \exp[-\theta y^\lambda]\, dy$$

$$= \underbrace{\left[-\exp[-\theta y^\lambda]\right]_0^\infty}_{F(y)} \quad \text{at } 0 = 1, \text{ at } +\infty = 0 \Rightarrow \text{equals 1.}$$

Probability of failure $= 1 - \exp[-\theta y^\lambda]$

Probability of survival $= 1 - F[y]$

Hazard function $h(y) = \frac{P(y|\lambda\theta)}{S(y)} = \lambda\theta y^{\lambda-1}$ (accelerated failure time) for weibull (function of y)

Exponential case is $\lambda = 1$ $h(y)$ is then θ (not a function of y).

independent of age "chance of survival same at 100 as 1000 etc"

lack of memory (y not important, may be a limitation)

2 Show it is a member of exponential distribution of family

$\lambda\theta y^{\lambda-1}\exp[\theta y^{\lambda}]$

$= \exp[-\theta y^{\lambda} + \log(\lambda\theta y^{\lambda-1})]$

$= \exp[-\theta y^{\lambda} + \log(\lambda) + \log(\theta) + \log(y^{\lambda-1})]$

$b(\theta)$ $a(y)$ $c(\theta)$ $d(y)$ λ is nuisance parameter

3 Expectation

$E[y] = \int_0^{\infty} y \cdot \lambda\theta y^{\lambda-1}\exp[\theta y^{\lambda}]\, dy$

Show

$E[y] = \left(\frac{1}{\theta}\right)^{\frac{1}{\lambda}} \Gamma\left[1+\frac{1}{\lambda}\right]$ with $\Gamma(u) = \int_0^{\infty} s^{u-1}\exp(-s)\, ds$

Substitute $u = \theta y^{\lambda}$ $du = \theta\lambda y^{\lambda-1}\, dy$

$= \int_0^{\infty} \lambda u \exp(-u)$

$E[y] = \int_0^{\infty} \left(\frac{u}{\theta}\right)^{1/\lambda}\exp(-u)$ look up proof online

look function applied to expectation

Check definition of weibull dist in R package, be careful

have to look back to get it WARNING

$\log\theta = -3.07 - 1.731x$ (the relationship) x is indicator for dist

To log re-scale β values

2.48 + 1267 (R number)

$\log\left(\frac{1}{\theta}\right)^{1/\lambda}$

08/03/16 APPLIED LINEAR STATISTICAL METHODS

MULTINOMIAL DISTRIBUTION

Extension to Binomial distribution

Binomial is a joint probability distribution

$$P(y_1, y_2, \dots y_J \mid \theta_1, \dots, \theta_J, n) = \frac{n!}{y_1! y_2! \dots y_J!} \theta_1^{y_1} \theta_2^{y_2} \dots \theta_J^{y_J}$$

$\vec{y} \begin{bmatrix} y_1 \\ \vdots \\ y_J \end{bmatrix}$ $\vec{\theta} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_J \end{bmatrix}$

$$p(\vec{y} \mid \vec{\theta}, n) = \frac{n!}{y_1! y_2! \dots}$$

can see response variable as a vector.

J : number of categories

For example: J = 3 yes, no, maybe

y_1 : # Say yes

y_2 : # Say maybe

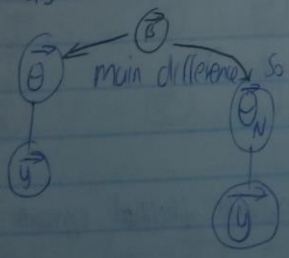
y_3 : # Say no

constraint: n : total number of answers $n = y_1 + y_2 + \dots + y_J$

in example $n = y_1 + y_2 + y_3$ for J = 3

constraint: $\theta_1 + \theta_2 + \dots + \theta_J = 1$

$E[y_1] = n\theta_1$ $E[y_2] = n\theta_2$ ...



So far, response variable a vector, NOT a scalar

Exercise 1: J = 2 verify that multinomial is binomial distribution

$$P(y_1, y_2 \mid \theta_1, \theta_2, n) = \frac{n!}{y_1! y_2!} \theta_1^{y_1} \theta_2^{y_2} \quad n = y_1 + y_2 \text{ and } \theta_1 + \theta_2 = 1$$

$$P(y_1 \mid \theta_1, n) = \frac{n!}{y_1!(n-y_1)!} \theta_1^{y_1} (1-\theta_1)^{n-y_1}$$ (binomial distribution)

Exercise 2 : is multinomial distribution a member of the exponential family of distributions ?

Exp : $p(y|\theta) = \exp[a(y).b(\theta) + c(\theta) + d(y)]$

$J=2$: yes, shown for binomial distribution

$J>2$: no

Even if multinomial distribution is not a member of the exponential family of distributions, we can control $\vec{\theta}\vec{y}$ collected over N groups via a set of parameters $\vec{\beta}$

NOMINAL LOGISTIC REGRESSION

$y=1$ yes
$y=2$ maybe
$y=3$ no

$$\log\left(\frac{\theta_j}{\theta_1}\right) = x^T\beta_j \qquad \beta \text{ depends on category } j.$$

J-1 equations

$$\begin{cases} \log\left(\frac{\theta_2}{\theta_1}\right) = x^T\beta_2 & \text{associated with "maybe" } J=2 \\ \qquad \rightarrow \text{using } \theta_1 \text{ as a reference} \leftarrow & \\ \log\left(\frac{\theta_3}{\theta_1}\right) = x^T\beta_3 & \beta\text{'s associated with "no" } J=3 \end{cases}$$

In general $\log\left(\frac{\theta_j}{\theta_1}\right) = x^T\beta_j \quad j = 2, \dots, J$

(θ_1 : reference category)

$\{\hat{\beta}_j\}_{j=2\dots J}$ are estimated with maximum likelihood approach

proportion associated with each category j : $\hat{\theta}_j = \frac{\exp(x^T\beta_j)}{1 + \sum_{j=2}^{J} \exp(x^T\beta_j)}$ $\quad \forall j = 2, \dots, J$

↖ SOFTMAX FUNCTION converting output to probability

$\hat{\theta}_j$ estimated from MLE

$\theta_j = \hat{\theta}_1 \exp(x^T\hat{\beta}_j)$ (thinking exp of both sides)

with $\theta_1 + \theta_2 + \dots \theta_J = 1$

08/03/16 APPLIED LINEAR STATISTICAL METHODS

Can rewrite as:

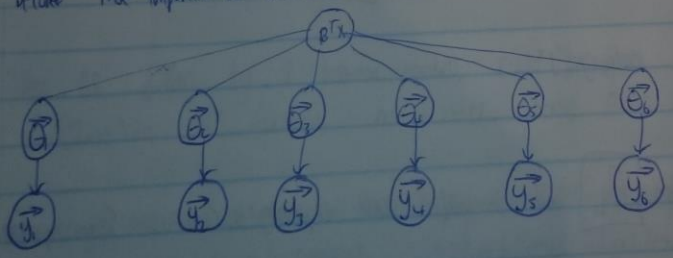
$$\hat{\theta}_1 + \hat{\theta}_1 \exp(x^T\hat{\beta}_2) + \ldots + \hat{\theta}_1 \exp(x^T\hat{\beta}_J) = 1$$

$$\hat{\theta}_1 = \frac{1}{1 + \exp(x^T\hat{\beta}_2) + \ldots + \exp(x^T\beta_J)}$$

EXAMPLE ON WEBSITE

Groups		y_1	y_2	y_3	n	GENDER	AGE or bel	Age2
[Lower 18-23]	$i=1$	$\vec{y}_1$ 26 26/45	12 12/45	7 7/45	$n_1 = 45$	F→0	0	0	0
[Lower 24-40]	$i=2$	$\vec{y}_2$ 96 96/245	210 210/245	15 15/245	$n_2 = 45$	F→0	1	1	0
[Lower >40]	$i=3$	$\vec{y}_3$ 5 5/60	14 14/60	41 41/60	$n_3 = 60$	F→0	2	0	1
[Low 18-23]	$i=4$	$\vec{y}_4$ 40 40/65	17 17/65	8 8/65	$n_4 = 65$	M→1	0	0	0
[Low 24-40]	$i=5$	$\vec{y}_5$ 17 17/44	15 15/44	12 12/44	$n_5 = 44$	M→1	1	1	0
[Low >40]	$i=6$	$\vec{y}_6$ 8 8/41	15 15/41	18 18/41	$n_6 = 41$	M→1	2	0	1
Saturated Model		θ_1	θ_2	θ_3				

Because MLE impossible for women without child.



indexed by "i"

$j = 2, 3$

Model 1: $\text{Log}\left(\frac{\theta_j}{\theta_1}\right) = \beta_{0j} + \beta_{1j}\,\text{Sex} + \beta_{3j}\,\text{Age}_1 + \beta_{4j}\,\text{Age}_2$

Model 2: $\text{Log}\left(\frac{\theta_j}{\theta_1}\right) = \beta_{0j} + \beta_{1j}\,\text{Sex} + \beta_{2j}\,\text{Age}$

model 4 has 2 less β's hence lower AIC.
model 2 has reduced complexity, not entirely a better fit.

What are $\hat{\theta}$'s of model 1?

-use Softmax function

	$\hat{\theta}_i$	i	$\hat{\theta}_5$
i=1	A	$\frac{\exp(\text{coef})}{1+\exp(\text{coef})+\exp(\text{coef})}$	$=0.24$ estimate coef for probability for first group
i=2			
i=3			
i=4	*	*	
i=5			
i=6			

$$\hat{\theta}_j = \frac{\exp(x^T \hat{\beta}_j)}{1 + \sum_{j=2}^{J} \exp(x^T \hat{\beta}_j)} \quad \text{SOFTMAX FUNCTION}$$

$$\hat{\theta}_{2i} = \frac{\exp(-0.59 - 0.38\text{Sex} + 1.13\text{Age1} + 1.59\text{Age2})}{1 + \exp(-0.59 - 0.38\text{Sex} + 1.13\text{Age1} + 1.59\text{Age2}) + \exp(-1.04 - 0.81\text{Sex} + 1.48\text{Age1} + 2.92\text{Age2})}$$

can do same for $\hat{\theta}_3$ and $\hat{\theta}_1$ is $[1 - \hat{\theta}_2 - \hat{\theta}_3]$

group "i" category "j"

Should be $\hat{\theta}_{2i}$ and i's under each exploring var i.e Sex$_i$ Age$_i$

Odd ratio: $\frac{\hat{\theta}_{2,1} / \hat{\theta}_{1,1}}{\hat{\theta}_{2,4} / \hat{\theta}_{1,4}} = 1.5$

1 ⇒ girls women 18-23

4 → men 18-23

men important/not v. important

Odd RATIO assuming Softmax function

Person of women of important over not important is bigger than that of men for age 18-23.

10/03/16 APPLIED LINEAR STATISTICAL METHODS

MULTINOMIAL DISTRIBUTION

Instead of one value of response, you have a vector of response (y)
Observe a number of groups with vector response, number of people in each group

$$\{(\vec{y}_i, x_i, n_i)\} \quad i = 1 \dots N \text{ (number of groups)}$$

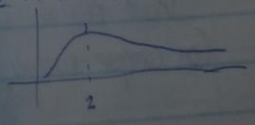
n_i: number of people voting in group i.
x_i: explanatory variable associated with group i.

J categories, $J-1$ degrees of freedom
number of betas by $J-1$
i.e. intercept, sex, age $= 3 \times 2 = 6$ d.f. in example.

d.f of saturated model: 2 d.f for each group (3 options per group)
6 groups = 12 d.f for saturated model

When AIC differs by 2 between models, could be from an extra variable
$(2m)$ in AIC formula, m is $m+1$ hence extra 2.

Comparing $m1$ and $m2$. $m1$ has 8 parameters, $m2$ has 6 parameters
The difference between log likelihood $-2[\text{Log}\mathcal{L}(m_1) - \text{Log}\mathcal{L}(m_2)] \approx 2$
Are the models that different?
$596 - 594 = 2$ from R output
$= 2 \sim \chi^2(\text{d.f } 2)$



2 will be in 95% CI of chi-squared.
Can now say $m1 \approx m_2$ basically equivalent.

ODD RATIO

i = group j = vote choice

$$OR = \frac{\theta_{j=2,i=1} / \theta_{j=1,i=1}}{\theta_{j=2,i=4} / \theta_{j=1,i=4}}$$

- ratio for group 1 women (18-23) [sex=0, age1=0, age2=0]
- ratio for group 4 men (18-23) [sex=1, age1=0, age2=0]

category 2 = y_2 $j=1$ = reference category

Only explanatory variable changing between groups is the sex

$OR = \exp[+0.388\ 130] \approx 1.5$ indicates the option is more important to women than to men

(used log as link function)

$$\log\left[\frac{\theta_{j=2,i=1}}{\theta_{j=1,i=1}}\right] = -0.59 - 0.38 sex_i + 1.12 Age1_i + 1.58 Age2_i$$

compute for $i=1$ and $i=4$ and compute ratio

only thing that will change is the $-0.38 sex_i$

How to use standard errors to ensure odds ratio outcome is correct?

$OR = \exp[+0.38 \pm 2 * 0.30]$ 0.30 standard error of sex in y_2

2*SE of sex (sex was being tested with J)

CI $[(\exp -0.38 - 2*0.3), (\exp -0.38 + 2*0.3)] = [0.8, 2.68]$ 95% CI on odds ratio

bc OR related to only one explanatory variable.

1 is within the CI ⇒ men + women may be having the same.

Compare group 3 to 6 because age2 or age1 will change. Will get new result. Should provide same result as previous test. end up with same β.

Can compare category 3 v category 1 (use y_3 ~~bet~~ β's). 1 will be outside this interval.

"not important" chosen as reference

5/23/16 APPLIED LINEAR STATISTICAL METHODS

Remark using Multinomial Distribution for nominal logistic Regression

Logit(θ_i) $\log\left[\frac{\theta_j}{\theta_1}\right] = X^T\beta_j \quad j=2,\dots,J$ J = # of categories (yes, no, maybe)

↑ ref category; Softmax function

Why log function?

When J=2: $\log\left[\frac{\theta_1}{\theta_2}\right]$ $\log\left[\frac{\theta_2}{\theta_1}\right] = X^T\beta_2$ only equation we have (binomial dist)

$\theta_1+\theta_2=1$

Alligators Example Handout

A Suited because it allows for multiple response categories. In this case 5 different categories

B Three pieces of information is enough to deduce which of four categories it belongs to. If L_1 L_2 and L_3 are all 0, then means it is Lake 4: Group

C $\log\left[\frac{\theta_j}{\theta_1}\right] = \beta_0 + \beta_1 Sex + \beta_2 Size + \beta_3 Lake_1 + \beta_4 Lake_2 + \beta_5 Lake_3 \quad j=2,3,4,5$

D For j=2,3,4,5

maximize $\log\left[\frac{\theta_j}{\theta_1}\right] = \beta_0 + \beta_1 Sex + \beta_2 Size + \beta_3 Lake_1 + \beta_4 Lake_2 + \beta_5 Lake_3$

E AIC: Akaike Information Criterion $= -2\log(L) + 2m$

L = likelihood function. m = number of parameters estimated

It is a measure of model fit, rewarding model accuracy and penalizing extra parameters

The lower the AIC, the better.

03/16

C

$$\theta_j = \frac{\exp(\beta_j^T x)}{1 + \sum_{k=2}^{J} \exp(\beta_k^T x)} \quad j = 2, \ldots, J \qquad \text{Softmax function}$$

$$\theta_1 = \frac{1}{1 + \sum_{k=2}^{J} \exp(\beta_k^T x)}$$

F. $\theta_2 = 0.17 - 0.46\text{Sex} - 1.34\text{Size} - 1.78\text{Lake}_1 + 0.91\text{Lake}_2 + 1.16\text{Lake}_3$

$\theta_3 = -3.42 - 0.63\text{Sex} + 0.56\text{Size} + 1.13\text{Lake}_1 + 2.53\text{Lake}_2 + 3.06\text{Lake}_3$

$\theta_4 = -2.43 - 0.61\text{Sex} + 0.73\text{Size} + 0.58\text{Lake}_1 - 0.55\text{Lake}_2 + 1.24\text{Lake}_3$

$\theta_5 = -1.43 - 0.25\text{Sex} - 0.3\text{Size} + 0.77\text{Lake}_1 + 0.03\text{Lake}_2 + 1.50\text{Lake}_3$

G.

SOLUTIONS

A. Multinomial distribution is an extension of the binomial distribution where the number of categories possible for the response is larger than 2. Here the number of categories is $J=5$, corresponding to 5 food choices available to alligators.

$$f(y_1, y_2, y_3, y_4, y_5 \mid \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, n) = \frac{n}{y_1! y_2! y_3! y_4! y_5!} \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3} \theta_4^{y_4} \theta_5^{y_5} \quad \text{with } y_1+y_2+y_3+y_4+y_5 = n$$

$$\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 = 1$$

B. Hancock: $(L_1, L_2, L_3) = (1, 0, 0)$ 4 different triplets to encode 4 level

Oklawaha: " " $= (0, 1, 0)$

Trafford: " $= (0, 0, 1)$

George: " $= (0, 0, 0)$

15/03/16 APPLIED LINEAR STATISTICAL METHODS

C $\text{Log}\left[\frac{\theta_j}{\theta_1}\right] = \beta_j^{\text{intercept}} + \beta_j^{\text{sex}}\,\text{sex} + \beta_j^{\text{size}}\,\text{size} + \beta_j^{L_1} L_1 + \beta_j^{L_2} L_2 + \beta_j^{L_3} L_3$

$\forall j = 2, \ldots, 5.$

$$\theta_j = \frac{\exp(\beta_j^T x)}{1 + \sum_{j=2}^{5} \exp(\beta_j^T x)} \qquad \forall j = 2, \ldots, 5 \text{ with } x = [1, \text{sex}, \text{size}, L_1, L_2, L_3]$$

$$\beta_j = [\beta_j^{\text{intercept}}, \beta_j^{\text{sex}}, \beta_j^{\text{size}}, \beta_j^{L_1}, \beta_j^{L_2}, \beta_j^{L_3}]$$

↑ $\exp(\beta_1^T x)$

$$\theta_1 = \frac{1}{1 + \sum_{j=2} \exp(\beta_j^T x)}$$

D $L = \prod_{i=1}^{16} f(y_{1i}, y_{2i}, y_{3i}, y_{4i}, y_{5i} \mid \theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}, \theta_{5i}, n_i)$ f = multinomial distribution

$n_i = y_{1i} + y_{2i} + y_{3i} + y_{4i} + y_{5i}$ group size can be different

Saturated likelihood of saturated model

$$\theta_{1i} = \frac{1}{1 + \sum_{j=2} \exp(\beta_j^T x_i)}$$ ← used to calculate fitted values in table

$$\theta_{ji} = \frac{\exp(\beta_j^T x_i)}{1 + \sum_{j=2} \exp(\beta_j^T x_i)}$$

$$L = \prod_{i=1}^{16} \frac{n_i}{y_{1i}!\, y_{2i}!\, y_{3i}!\, y_{4i}!\, y_{5i}!} \theta_{1i}^{y_{1i}} \theta_{2i}^{y_{2i}} \theta_{3i}^{y_{3i}} \theta_{4i}^{y_{4i}} \theta_{5i}^{y_{5i}} \qquad d.f = 4 \times 16 - y_i$$

↑ ↑

$L(\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}, \theta_{5i}) = \; i = 1 \ldots 16 \qquad d.f = 6 \times 4 - \beta_{ij}$

$L(\{\beta_j\}_{j=2 \ldots 6})$

E Criterion to use to select best model (can use to drop parameters or not)

03/16

F $\beta_0^T x = 0.17 - 0.465 sx - 1.36 size - 1.86 L_1 + 0.41 L_2 + 1.16 L_3$

$\beta_1^T x = -7.42 - 0.635 sx + 0.565 size + 1.134 + 2.532 L_2 + 3.06 L_3$

$\beta_2^T x = -2.03 - 0.115 sx + 0.735 size + 0.534 - 0.554 L_2 + 1.24 L_3$

$\beta_3^T x = -1.43 - 0.235 sx - 0.245 size + 0.774 + 0.036 L_2 + 1.56 L_3$

G the table of fitted values

	Hancock			vs	Oh Iowaches			Use table 8. Odd ratio
one Small	(1) $i=1$	$\frac{0.05}{0.6}$	bad Fish 0.083		$i=5$	$\frac{0.08}{0.30}$	bad Fish 0.026	3.08
male Big	$i=2$	$\frac{0.11}{0.62}$			$i=6$	$\frac{0.028}{0.48}$		3.08
female Small	$i=3$	$\frac{0.049}{0.53}$			$i=7$	$\frac{0.01}{0.21}$		3.08
female Big	$i=4$	$\frac{0.13}{0.52}$			$i=8$	$\frac{0.058}{0.36}$		3.08

$OR = \exp\left(\beta_{j=4}^{L_1} - \beta_{j=4}^{L_2}\right)$

$0.58 - (-0.55) = 1.3$

H Confidint? SE: $\beta_{j=4}^{L_1} = 0.79$ $\beta_{j=4}^{L_2}$ 1.20

take Sqrt $SE\left[\beta_{j=4}^{L_1} - \beta_{j=4}^{L_2}\right]^2 = \left[SE\left(\beta_{j=4}^{L_1}\right)\right]^2 + \left[SE\left(\beta_{j=4}^{L_2}\right)\right]^2$ independance assumed

$SE = \sqrt{0.79^2 + 1.20^2} = 1.45$

$1.13 \pm 1.45 * 2$ = 95% CI 0 is in interval. Not confident.

exp(0)=1, use 0 since that OR can or will be 1.

can't say for sure "allegator ... in Oh ... less than Hancock"

22/03/16 APPLIED LINEAR STATISTICAL METHODS

INFERENCE

Algorithms to estimate β's

Usually $\hat{\theta} = \arg\max_{\theta} lik(\theta)$

$= \arg\max_{\theta} \log(lik(\theta))$ usually differentiate this function and equate to 0.

Normal Distribution

link g: identity function [mapping $\mathbb{R}$ onto $\mathbb{R}$]

$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(y-\theta)^2}{2\sigma^2}\right]$

$y_1, \ldots, y_n$: N responses

$x_1, \ldots, x_n$

$x_i^T\beta$ (θ_i constrained by data) (constraining the mean)

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

$\log(lik(\beta)) = -N\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{N} (y_i - x_i^T\beta)/2\sigma^2$

$SSE = \sum_{i=1}^{N} (y_i - x_i^T\beta)^2$

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

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

$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & & & \vdots \\ 1 & x_{1N} & & x_{pN} \end{bmatrix}$ i-th row

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

$= \frac{d}{d\beta}\left[\vec{y}^T\vec{y} - \vec{y}^T X\beta - (X\beta)^T\vec{y} + (X\beta)^T X\beta\right]$ equate to zero

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

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

1. Differentiate log likelihood ∇loglik
2. Set $\nabla \text{loglik} = 0$
3. Solving system of equations defined by $\nabla \text{loglik} = 0$

Newton - Raphson

Looking at $\text{loglik}(\theta)$ near $\theta^{(i)}$ — Taylor expansion

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

$\nabla \text{loglik}(\theta^{(i)})$: gradient at $\theta^{(i)}$

$\nabla\nabla^T \text{loglik}(\theta^{(i)})$: Hessian matrix at $\theta^{(i)}$: $H_{\theta^{(i)}}$

After differentiating: $\nabla \text{loglik}(\theta) \simeq H_{\theta^{(i)}}(\theta - \theta^{(i)}) + \nabla \text{loglik}(\theta^{(i)}) = 0$

Newton Raphson method:

Initialize $\theta^{(i)} = \theta^{(0)}$

$$\theta^{(i+1)} = \theta^{(i)} - [H_{\theta^{(i)}}]^{-1} \nabla \text{loglik}(\theta_i)$$

(inverse × gradient)

until convergence (J)

If loglike is quadratic, NR will find solution in one step (i.e. normal distribution).
In GLM, 1 or one extrema, no local extremas.

The quantity $[-H]$ (Hessian) (Observed information) determines the sharpness of the peak in the likelihood function around its maximum

Alternative to Newton Raphson: METHOD OF SCORING

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

Replaces J by $I^{\vee}$ in NR method

20/09/16 ALSM 2 EXAM NOTES

Binomial

$P(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$ $y \in (0, 1, \dots, n)$ Logit $\left(\frac{p}{1-p}\right)$ Prob $\in$ -Inf to +Inf

Distribution: $= [p + (1-p)]^n = [p]^n = 1$ by binomial expansion

Exponential: $= \exp\left[y\log\left(\frac{\theta}{1-\theta}\right) + \log\binom{n}{y} + n\log(1-\theta)\right]$

Maximum: $\left[\frac{\partial}{\partial\theta}\right]\left[\theta^y(1-\theta)^{n-y}\right]\left[y(1-\theta)^{-1}(n-y)\right]$ $\theta = y/n$

Expected Value: $y\binom{n}{y} = n\binom{n-1}{y-1}$ $= n\theta(p+1-p)^{n-1} = n\theta$

POISSON DISTRIBUTION

$P(y|\theta) = \frac{\lambda^y}{y!}\exp[-\lambda]$ $\lambda > 0$ $y \in \mathbb{N}$

Distribution: $\exp(-\lambda) \leq \frac{\lambda^y}{y!}$ Taylor expansion $\exp(-\lambda)\exp(\lambda) = 1$

Exponential: $\exp[y\log(\lambda) - \log(y!) - \lambda]$

Expectation: $\exp(-\lambda \leq \frac{y\lambda^y}{y!} = " \leq \frac{\lambda^{y-1}y}{y(y-1)!}$ $= \lambda\exp(-\lambda)\exp(\lambda) = \lambda$

Maximum: $\left[\frac{\lambda^y\exp(-\lambda)}{y!}\right][y-\lambda] = 0$ $\lambda = y$

Exponential DISTRIBUTION

$P(y|\theta) = \theta\exp(-\theta y)$ $y \in \mathbb{R}^+$ $\theta \in \mathbb{R}^{+*}$

Distribution: $\int_0^\infty \theta\exp(-\theta y) = -[\exp(-\theta y)]_0^\infty$ $= 0+1 = 1$

Exponential: $\exp[\log(\lambda) - \lambda y]$

Expectation: $[-y\exp(-\lambda y)]_0^\infty \int_0^\infty \exp(-\lambda y)dy$ $[0-0] + [\frac{-1}{\lambda}\exp(-\lambda x)]_0^\infty = \frac{1}{\lambda}$

Maximum: $\exp[-\lambda y] - \lambda y\exp(-\lambda y)$ $\exp[-\lambda y][1-\lambda y] = 0$ $\lambda = \frac{1}{y}$ (mean)

Failure F(y): $\int_0^y \lambda\exp(-\lambda y)dy = [-\exp(-\lambda y)]_0^y$ $= 1 - \exp(-\lambda y)$

Survival S(y): $1 - F(y)$ $= \exp(-\lambda y)$

Hazard: original function/survival fun $\frac{\lambda\exp(-\lambda y)}{\exp[-\lambda y]} = \lambda$ [memoryless, not dependent on time]

Weibull DISTRIBUTION

$P(y|\lambda,\theta) = \lambda\theta y^{\lambda-1}\exp[-\theta y^\lambda]$ $y \in \mathbb{R}^+$ $\theta \in \mathbb{R}^{+*}$ $\lambda \in \mathbb{R}^{+*}$

Distribution: $\int_0^\infty$ $= [-\exp(-\theta y^\lambda)]_0^\infty$ $= 1+0 = 1$

Exponential: ? $\exp(-\theta y^\lambda + \log(\lambda\theta y^{\lambda-1})$ $= \exp[-\theta y^\lambda + \log(\lambda) + \log(\theta) + (\lambda-1)\log(y)]$

Expectation: $\left(\frac{1}{\theta}\right)^{1/\lambda}\Gamma\left(1+\frac{1}{\lambda}\right)$

Maximum: ?

Failure: $\int_0^y p(y|\theta,\lambda)$ $= 1 - \exp[-\theta y^\lambda]$ Hazard: $\frac{p(y|\theta,\lambda)}{s(y)}$ $= \lambda\theta y^{\lambda-1}$ (accelerated Failure law, function of y)

Survival S(y): $1 - F(y)$ $= \exp[-\theta y^\lambda]$

Multi-Nomial Distribution

$P(y_1, y_2 | \theta_1, \theta_2, n) = \frac{n!}{y_1! y_2!} \theta_1^{y_1} \theta_2^{y_2}$

- Generalization of binomial. Binomial when $J=2$. When $J \geq 2$ not member of exponential family
- J = categories. θ_i probability of category, $y_1 + y_2 + \dots + y_J = n$
- We can write $\vec{y}$ collected over N groups via a set of parameter $\vec{\beta}$
- Reference category selection eg θ_1

$\text{Log}(\theta_j) = [\theta_j / \theta_1] = x^T \beta_j \quad \forall j = 2, \dots, J$ having constraint $\sum_{j=1}^{J} \theta_j = 1$

- When parameters $\vec{\beta}$ are estimated then $\theta_j = \theta_1 \exp(x^T \beta_j) \quad \forall j = 2, \dots, J$

$\theta_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp(x^T \beta_j)}$

$\theta_j = \frac{\exp(x^T \beta_j)}{1 + \sum_{j=2}^{J} \exp(x^T \beta_j)} \quad \forall j = 2, \dots, J$ SOFTMAX FUNCTION ⇒ converting output to probability

Intercept $\times$ (category -1) D.F.

Saturated model: Number groups $\times$ (# Option -1)

Alligator example

Saturated model has $N(J-1)$ D.F.

20/05/16 ALSM 2 EXAM NOTES

BINOMIAL AND POISSON GLMS

$H_0: \theta \leq 0.5$ v $H_1: \theta > 0.5$ if $p < \alpha$ accept H_1

$p = \frac{\int_{0.5}^{1} \binom{n}{y} \theta^y (1-\theta)^{n-y} \, d\theta}{\int_{0}^{0.5} \binom{n}{y} \theta^y (1-\theta)^{n-y} \, d\theta + \int_{0.5}^{1} \binom{n}{y} \theta^y (1-\theta)^{n-y} \, d\theta}$

DEVIANCE

- Null deviance: $-2 \log \left[\frac{\text{likelihood of null model (1 parameter)}}{\text{likelihood of saturated model}} \right]$
- Residual Deviance: $-2 \log \left[\frac{\text{likelihood of current model}}{\text{likelihood of saturated model}} \right]$
- Measure of goodness of fit, smaller values better
- If model is good deviance follows χ^2 distribution with $n-m$ D.F. and 0 D.F.
- If $D_{model} < \chi^2_{n-m,\alpha}$ then it is a good model

AIC

$AIC = -2 \log L(\hat{\beta}) + 2p$

- $p = \dim(\beta)$ = number of estimated parameters
- $\hat{\beta}$ are the estimated parameters that maximise the likelihood of the model
- $\log L(\hat{\beta})$ is maximum value of log likelihood
- Best model is a tradeoff between the maximised likelihood with minimum number of parameters
- Select model with smallest AIC

SURVIVAL ANALYSIS

- Time until failure $y \in \mathbb{R}^+$
- Probability of failure: $\int_0^y p(y|\theta) \, dy$ $F(y)$
- Probability of survival: $1 - F(y)$
- Hazard function $\frac{P(y \leq t \leq y+\delta y)}{P(y \leq t)}$ Chance of failure between the year y_t / Surviving beyond time y $S(y)$
- $h(y) = p(y|\theta)/S(y)$
- "Chances of dying in the next minute, given a survival rate and given you have not died before time y"
- Cumulative hazard function $H(y) = -\log [S(y)]$
- When $F(y) = S(y) = 0.5$ called the "median point of failure"
- When y_i is censored, use $S(y)$ instead of $p(y_i|\theta_i)$

: $\prod_{i=1}^{N} p(y_i|\theta)^{\delta_i} \, S(y_i)^{1-\delta_i}$

NEWTON RAPHSON

Initialize $\theta^{(1)} = \theta^{(0)}$

$\theta^{(t+1)} = \theta^{(t)} - [H_{\theta^{(t)}}]^{-1} \nabla \ell(\theta^{(t)})$

Until Convergence — only one extension in GLM (no local extrema)

Method Of Scoring

$I = E[s|\theta]$ - Expected Fisher Information — expectation of Hessian taken wrt. θ

Replace $-[H_{\theta^{(t)}}]^{-1}$ by I in NR method

Limitation of Method

- Assumed responses are independent
- Outliers can give a zero all data except one group vote for parameter β

GLM STEPS ✓

- Bernoulli ✓
- Binomial ✓
- Poisson ✓
- link between Poisson/Binomial ✓
- Normal / Gaussian
- Multinomial
- Survival: Weibull
  Exponential
- Deviance / AIC ✓
- Binomial Poisson ✓

1. Show this is a distribution
2. Show member of exponential family
3. Compute expectation of y, function of θ
4. Compute to maximum with θ

Basic SLR ✓

ANOVA DECOMPOSITION ✓ DF (S) ✓

Properties Expectation and Variance ✓

Hypothesis testing / Confidence Intervals ✓

F-statistic - P value ✓

CI and PI ✓

Problem Sheet 1 ✓
2 ✓
3
4

Matrix Formulation of SLR Model

Multiple Regression

ANOVA

CI PI

Sequential and Partial SS

General linear hypothesis ($L\beta = m_0 L$)

Orthogonality / Multicollinearity?