

Lecture 1/19

Friday, December 29, 2017 9:55 AM

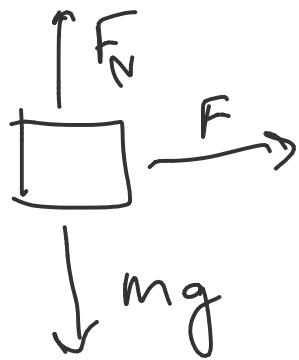
- We will use Blackboard for this class.
 - Each problem set has data on the Website.
 - Office Hours: MW 2:15-3:15
R 11 - 12
 - Texts: McCullagh - mostly classical
Faraway - mostly applied
McCullagh - mostly applied
- Grading - Biweekly problem sets 40%
- Theory & Applied
- Exams 40%
- Take home midterm
- In class final

Intro to GLM

Goal: Exam relationships between variables (X, Y).

- LM provides one approach to modeling this relationship

Ex:

$$\begin{array}{c} F_N \\ \square \\ mg \end{array} \rightarrow \quad a = F/m$$


Why linear, Newton? Empirically.

$$\text{Stochastic LM: } a_i = \frac{F_i}{m} + \varepsilon_i.$$

Start to think about C.I.

and quantifying the uncertainty
in the model.

- Conditioning can really affect our intuition about which line is best
 - Minimization goal/error classification matters.

Classical Problems

- Continuous response/predictor
- Linear relationship
- ε_i are symmetric
- Linear in "linear models" is the connection between Y and the parameters
- We can even fit

$$y_i = \alpha \cos(x_i - b) + \varepsilon_i$$

$$= c \cos(x_i) + d \sin(x_i) + \varepsilon_i$$

Now we can use OLS estimates
for (c, d)

- Predictors need not be continuous.

$$y_i = \begin{cases} m_1 + \varepsilon_i \\ m_2 + \varepsilon_i \end{cases}$$

$$= m_i + \alpha I_2 + \varepsilon_i$$

then inference between group 1 and 2 is just inference on α .

Q: Where does linear models fail?

Ex: $y_i = \begin{cases} 0 \\ 1 \end{cases}, y_i = \text{count}$

Y_i = service time

All examples where $Y_i \not\sim N(\mu, \sigma^2)$

Topics

- Extend the linear model where the data is not

Gaussian.

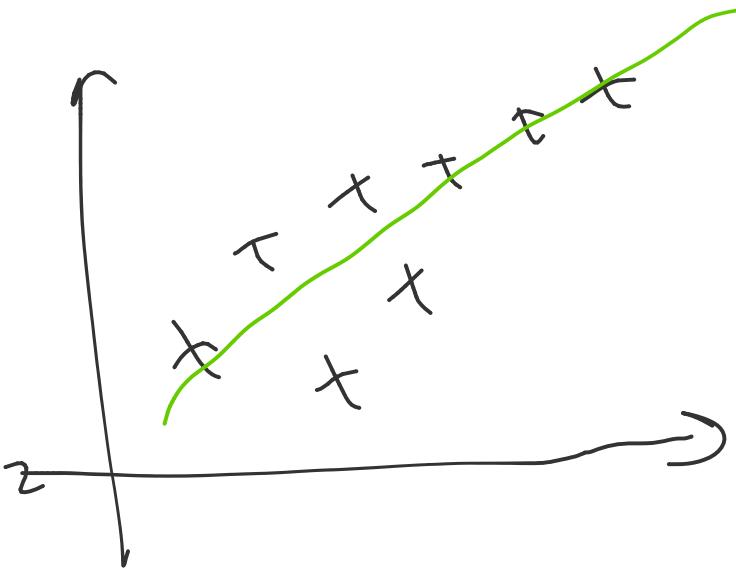
Outline

This time: linear model
review

Reading: LM review

Linear Model Review

Let $\{(x_i, y_i)\}_{i=1}^n$ be
a collection of data.



A linear regression model
is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

deterministic stochastic

X is our predictor or independent variable.

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

independent

is the model parameter

{

zero mean R.V.

(usually assumed)

$$\varepsilon_i \sim N(0, \sigma^2)$$

y is called the response

or dependent variable

Having ε be zero

mean is really restriction.

Equivalently we could
written

I have written

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

or

$$f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - (\beta_0 + \beta_1 x))^2}{2\sigma^2}\right\}$$

Assuming that $(Y_i)_{i=1}^n$

are independent samples
from this conditional
model

$$\begin{aligned} f_{X_1, \dots, X_n | X_1 = x_1, \dots, X_n = x_n}(y_1, \dots, y_n) \\ = \prod_{i=1}^n f_{Y_i | X_i = x_i}(y_i) \end{aligned}$$

$$= \prod_{i=1}^n f_{Y_i|X_i=x}(y)$$

$$= L(\beta_0, \beta_1)$$

$$= \exp \left\{ - \sum_{i=1}^n (y_i - (\beta_0 + \beta_1))^2 - C \right\}$$

$$\text{for } C = \frac{-n}{2} \log(2\pi\sigma^2)$$

Goal: Select parameters

$(\hat{\beta}_0, \hat{\beta}_1)$ that are "consistent" with the data.

Approaches:

- Minimizing

- $\arg \min$

some cost

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

- $\underset{(\beta_0, \beta_1)}{\operatorname{argmin}} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$
- Maximizing Likelihood (or $\log L$)
- $\underset{(\beta_0, \beta_1)}{\operatorname{argmax}} - \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$
- \Leftrightarrow
- $\underset{(\beta_0, \beta_1)}{\operatorname{argmin}} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$
- This Shows
 $\hat{\beta}_{OLS} = \hat{\beta}_{MLE}$
- So really, OLS estimates
 really comes from MLE.
 ... turns

- In this class, we focus on Likelihood based estimation techniques.

- Finding the estimators

$$\frac{\partial}{\partial \beta_0} \log L = -2 \sum_{i=1}^n \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2} = 0$$

$$\Rightarrow \sum y_i = n \hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

$$\Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^n \frac{2(y_i - (\beta_0 + \beta_1 x_i))}{2\sigma^2} x_i = 0$$

~ ~ ~ 2

$$\partial \hat{\beta}_1$$

$$\Rightarrow \sum y_i x_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

$$\sum y_i x_i = \bar{y} \sum x_i - \hat{\beta}_1 \bar{x} \sum x_i + \hat{\beta}_1 \sum x_i^2$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum y_i x_i - \bar{y} \sum x_i}{\sum x_i^2 - \bar{x} \sum x_i}$$

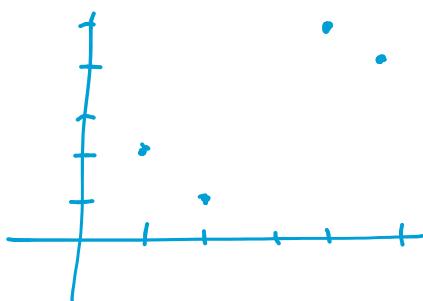
$$= \frac{\sum (y_i x_i - \bar{y} x_i)}{\sum (x_i^2 - \bar{x} x_i)} = \frac{s_{xy}}{s_{xx}}$$

$$= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

Turns out: These are UMVUE.

Ex:

X	1	2	4	5
Y	2	1	5	4



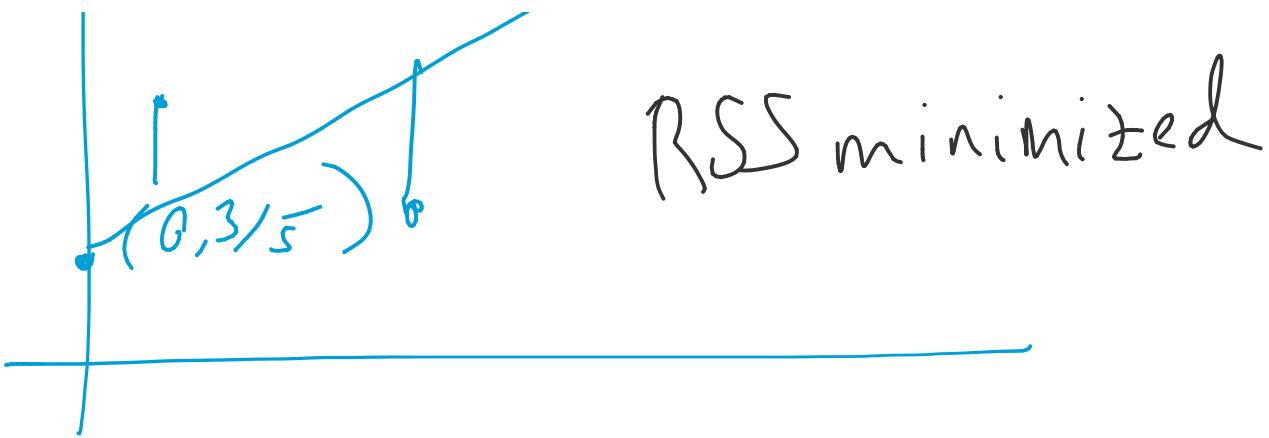
$$\bar{y} = \bar{x} = 3$$

$$\hat{\beta}_1 = \frac{(-2)(-1) + (-1)(-2) + (1)(2) + (2)(1)}{2^2 + 1^2 + 1^2 + 2^2}$$

$$= \frac{8}{10} = \frac{4}{5}$$

$$\hat{\beta}_0 = 3 - \left(\frac{4}{5}\right)(3) = \frac{3}{5}$$





Vector Version SLR

Let $\{(x_{i1}, x_{i2}, \dots, x_{ip}, y_i)\}_{i=1}^n$

be a collection of data.

Let $x_i = (x_{i1}, \dots, x_{ip})^T$

then the design matrix

is given by

$$X = [1 | x_1 | x_2 | \dots | x_{i1} | \dots | x_p]$$

$$X = \begin{bmatrix} 1 & x_1 & \dots & x_p \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$$

$$\text{for } 1 = (1, \dots, 1)^T \in \mathbb{R}^{n \times 1}$$

Then the linear regression model can be written as

$$Y = X\beta + \varepsilon \quad \text{for}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{and} \quad \varepsilon \sim MVN(0, \sigma^2 I)$$

Or equivalently

$$Y | X \sim MVN(X\beta, \sigma^2 I)$$

$$Y|X \sim MVN(\Lambda(\beta_r^0 -))$$

Overview

Last Time: Scalar LM

This Time: Vector LM
+
Inference

Reading: MAN Chp 1, 2.1 - 2.2

Vector Linear Models

SLR (vector): $\mathbf{Y} = \mathbf{X}\beta + \epsilon$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In classical linear models

we assume

$$\vec{\varepsilon} \sim MVN(\vec{0}, \sigma^2 I)$$

Goal: Estimate β

by (1) LS: $\underset{b}{\operatorname{argmin}} (Y - X\beta)^T(Y - X\beta)$

(2) MLE: $Y|X \sim MVN(X\beta, \sigma^2 I)$

which maximizes

$$-\frac{1}{2} \sum_{i=1}^n \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2$$

$$I(\beta) = \left(\frac{1}{\sqrt{2\pi}\sigma^2} \right)^n \exp \left(-\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta) \right)$$

$$= \exp \left\{ -\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta) - \frac{n}{2} \log [2\pi\sigma^2] \right\}$$

Goal is to now find

$$\arg \max_{\beta} I(\beta) = \arg \min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

This again shows that

$$\hat{\beta}_{MLE} = \hat{\beta}_{OLS}$$

Now finding $\hat{\beta} = \hat{\beta}_{MLE} = \hat{\beta}_{OLS}$

$$\text{Max } L \quad \gamma, \tau, P, V, \nu, r \stackrel{\text{set}}{=} \cap$$

$$\frac{\partial \text{log } L}{\partial \beta^T} = 2X^T(Y - X\beta) \stackrel{\text{set}}{=} 0$$
$$\Rightarrow X^T Y - X^T X \beta = 0$$

$$X^T Y = X^T X \beta$$

Now if $(X^T X)^{-1}$ exists

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

If $(X^T X)^{-1}$ doesn't exist
then there isn't one
unique solution to this
equation.

equation.

We sometimes say the model is not identifiable.

Possible Solutions

(A) Use generalized inverse / regularization

(B) Eliminate redundant predictors.

• $B \subseteq A$ because we just require $\beta_j = 0$ for $j \in J$ labeling the redundant

using the values
indices.

Ex: $X = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$ $y = \begin{bmatrix} 2 \\ 1 \\ 5 \\ 4 \end{bmatrix}$

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

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 12 \\ 12 & 46 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{184-194} \begin{bmatrix} 4 & -12 \\ -12 & 4 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 12 \\ 4 & 4 \end{bmatrix}$$

$$\hat{\beta} = \frac{1}{40} \begin{bmatrix} 4 & -12 \\ -12 & 4 \end{bmatrix} \begin{bmatrix} 12 \\ 4 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix} \text{ which matches the scalar version.}$$

Inference in LM

Properties of $\hat{\beta}$:

$$\mathbb{E}(\hat{\beta}|x) = \mathbb{E}((x^T x)^{-1} x^T y | x)$$

$$= (x^T x)^{-1} x^T \mathbb{E}(y|x)$$

$$= (x^T x)^{-1} x^T x \beta$$

$$\rightarrow \beta$$

So $\hat{\beta}$ is unbiased.

$$\text{Var}(\hat{\beta}|x) = (x^T x)^{-1} x^T \text{Var}(y|x) x (x^T x)^{-1}$$

$$= (x^T x)^{-1} x^T \sigma^2 I x (x^T x)^{-1}$$

$$= \sigma^2 (x^T x)^{-1} x^T x (x^T x)^{-1}$$

$$= \sigma^2 (x^T x)^{-1}$$

$$= \sigma^2(X'X)^{-1}$$

We show in HW that
 $\hat{\beta}$ is an efficient, UMVUE,
BLUE estimator.

We also know its true dist.

$$\hat{\beta} \sim MVN(\beta, \sigma^2(X'X)^{-1})$$

If σ^2 is unknown and n is not sufficiently large.

first obtain estimate $\hat{\sigma}^2$

$$\text{by } \hat{\sigma}^2 = \frac{1}{n} (Y - X\hat{\beta})(Y - X\hat{\beta})'$$

which is the maximum likelihood estimate of σ^2 .

- Turns out this is the minimum MSE predictor.
- But we can correct to make it unbiased.

Properties:

- $n \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$

- $\hat{\beta}_i \sim t_{n-(p+1)}$

$$\frac{t_{\alpha/2}}{\sqrt{\text{Var}(\hat{\beta}_i)}} \sim t_{n-(p+1)}$$

for $\text{Var}(\hat{\beta}_i) = \text{diag}_i(\sigma^2(X^T X))$

which gives the

$(1-\alpha)$ CI for $\hat{\beta}_i$

$$\hat{\beta}_i \pm t_{\alpha/2, n-(p+1)} \sqrt{\text{Var}(\hat{\beta}_i)}$$

Ex:

$$(X^T X)^{-1} = \frac{1}{40} \begin{bmatrix} 46 & -12 \\ -12 & 4 \end{bmatrix}$$

$$\hat{\theta}^2 = \frac{1}{4} \left(\left(\frac{\beta}{5}\right)^2 + \left(\frac{\epsilon}{5}\right)^2 + \left(\frac{\zeta}{5}\right)^2 + \left(\frac{\gamma}{5}\right)^2 \right)$$

$$= 9/10$$

$$\text{Var}(\hat{\beta}_1) = \frac{9}{100}$$

$$\text{se}(\hat{\beta}_1) = \frac{3}{10}$$

So then a 95% CI is given by

$$0.8 \pm 4.3(3/10)$$

$$\therefore = (-1/2, 2.1)$$

Overview

Last time: Inference in
Standard linear model

This Time: Exponential

Family of distributions

& defining a GLM

Reading: M&N 2.3-2.5

Linear Models Beyond Normal.

D_{of}: A distribution $f(y; \theta)$

Def: A distribution $t(y)$ is said to be in the exponential family iff

$$f(y; \theta) = \exp \left\{ w(\theta) t(y) + b(\theta) + c(y) \right\}$$

for $y \in A$ where

$w(\theta)$ & $b(\theta)$ don't depend on y
 $t(y), c(y)$, & A don't depend on θ

- Most common distributions are in the exponential

- family
- o Normal |
 - o Exponential |
 - o Γ, β, χ^2
 - o Binomial |
 - o Poisson

Not in exp family

- o Unif(a, b)

Ex: Show that $\mathcal{Y} \sim N(\mu, \sigma^2)$
with μ unknown σ^2
known.

$$f(y; \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

$$= \exp\left\{\underbrace{\frac{y\mu}{\sigma^2} - \frac{y^2}{2\sigma^2}}_{w(\mu)t(y)} - \underbrace{\frac{\mu^2}{2\sigma^2}}_{c(y)} - \underbrace{\frac{1}{2}\ln(2\pi\sigma^2)}_{b(\mu)}\right\}$$

Ex: Show that $Y \sim \text{Pois}(\lambda)$ is in the exp. family.

$$f(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \quad A = \{0, 1, \dots\}$$

$$= \exp\left\{-\lambda + \underbrace{y\ln\lambda - \ln y!}_{t(\lambda)}\right\}$$

$$1(\lambda) \quad t_1(\lambda) \quad c(y)$$

$$b(\lambda) \quad \underline{w(\lambda)} \quad t(y) \quad c(y)$$

If we let $\eta = w(0)$
then if w^{-1} exist.

$$f(y; \eta) = \exp\left\{\eta t(y) + b^*(y) + c(y)\right\}$$

We say that this dist.
is in canonical form

and η is the canonical
parameter.

For iid data

$$l(\eta) = \exp \left\{ \eta \sum t(y_i) + b^*(\eta) + \{c y_i\} \right\}$$

So

$\sum t(y_i)$ is a sufficient statistic for η and by extension θ .

In M&N, they develop theory based on a particular form of the exp family.

$$f(y; \theta, \phi) = \exp \left\{ \frac{y^\theta - b(\theta)}{\alpha(\phi)} + c(y, \phi) \right\}$$

for $y \in A$ ind of θ .

How is this different?

- Already in canonical form
- y has been defined so that the sufficient statistic is $\sum y_i$
- ϕ is called the scale or dispersion parameter.
- For the time being ϕ

is known.

Ex: Normal: $\theta = \mu$, $a(\phi) = \sigma^2$

$$b(\theta) = \frac{\theta^2}{2}$$

Ex: Poisson: $\theta = \log \lambda$, $a(\theta) = 1$

$$b(\theta) = e^\theta$$

(carefull with how we
define θ).

All dist the exp family
can be written in this
natural exp. family form.

Setting up a GLM:

Assume $(Y_i)_{i=1}^n$ are

obs. from the exp.

family $Y_i \sim f(y; \theta, \phi)$

Let $\mu_i(\theta) = \mathbb{E}[Y_i]$.

Define a link function

$g(\cdot)$ [invertible] and

define $\eta_i = g(\mu_i)$.

We construct a GLM

by setting

by setting

$$Y_i = \vec{x}_i \beta^T$$

where \vec{x}_i is the design vector of covariates

and β is a vector

of fixed, unknown parameters.

Choosing a Link

A useful choice: the canonical link function is the function g s.t.

$$\mu_i = \underline{\theta_i}$$

Canonical parameter.

Ex: (Normal)

$$g(\mu_i) = \theta = \mu_i$$

So g is the identity function/link

$$\text{Set } \mu_i = \vec{x}_i^T \vec{\beta}$$

Ex: (Poisson)

$$g(\mu_i) = \log(\mu_i) = \log(\lambda_i)$$

So $g()$ is the log link.
r + T.

Set. $\log(\lambda_i) = \vec{x}_i^T \vec{\beta}$

$$\lambda_i = e^{\vec{x}_i^T \vec{\beta}}$$

Overview

Last Time: EXP families
and def. of a GLM

This Time: Mean/Variance
functions, ML Estimation

Reading: M&N Chp 2.3-2.5

Notes: PS #1 due soon

Project datasets online.

Natural Exponential Family (NEF).

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{\sum x_i \theta_i - b(\theta_i)}{a(\phi)} + (y_i, \phi) \right\}$$

GLM: For $y_i \in \text{NEF}$ and

a link function $g(\cdot)$

Set

$$\eta_i = g(\mu_i) = \vec{x}_i^T \beta_i$$

Question: How can we compute
 \dots

QUESTION: " ... "

μ_i for any $f \in \text{NEF}$.

Well,

$$\mu_i = \int y_i f(y_i) dy_i$$

This could be hard...
Use this trick instead.

$$\int_A f(y_i; \theta_i) dy_i = 1$$

So

$$\frac{\partial}{\partial \theta_i} \int f(y_i; \theta_i) dy_i = 0$$

A does not depend on θ
 So we can use Leibnitz to
 interchange.

$$\Rightarrow \int_A \frac{\partial}{\partial \theta_i} \exp \left(\frac{y_i \cdot \theta_i - b(\theta_i)}{a(\phi)} + c \right) dy_i = 0$$

$$\Rightarrow \int_A \left(\frac{y_i - b'(\theta_i)}{a(\phi)} \right) f(y_i; \theta_i) dy_i = 0$$

$$\Rightarrow \mathbb{E}[y_i - b'(\theta_i)] = 0$$

$$\Rightarrow \mathbb{E}[y_i] = b'(\theta_i)$$

$$\Rightarrow \mu_i = b'(\theta_i)$$

To find its variance

$$\text{Var}(Y_i) = V_i$$

$$O = \int_A f(y_i; \theta_i) \left[\frac{(y_i - b'(\theta_i))^2}{a(\phi)} - b''(\theta_i) \right] dy_i$$

$$\int_A (y_i - \mu_i)^2 f(y_i; \theta_i) dy_i = b''(\theta_i) a(\phi) \int_A f dy_i$$

$$V_i = b''(\theta_i) a(\phi)$$

H

$$V_i = b''(\theta_i) a(\phi)$$

A

For this reason we
say $b(\cdot)$ is the
cumulant function.

Ex: $N(\mu, \sigma^2)$

unknown known

$$f(y_i; \mu) = \exp \left\{ \frac{y_i - \frac{\mu}{2}}{\sigma^2} + c \right\}$$

then

$$a(\phi) = \sigma^2 \quad b(\mu) = \mu^2 / 2$$

$$a(\phi) = \theta^* \quad b(\mu) = \mu/2$$

$$\mathbb{E}(Y) = b'(\mu) = \mu$$

$$\text{Var}(Y) = \theta^2 b''(\mu) = \theta^2 \cdot 1$$

Ex: (Poisson)

$$f(Y; \lambda) = \exp \left\{ \frac{Y \cdot \log \lambda - \lambda}{1} - C \right\}$$

$$= \exp \left\{ \frac{Y\theta - e^\theta}{1} - C \right\}$$

for $\theta = \log \lambda$

and have

$$a(\theta) = 1$$

$$b(\theta) = e^{\theta}$$

$$\mathbb{E}(Y) = b'(\theta) = e^{\theta} = \lambda$$

$$\text{Var}(Y) = 1 \cdot b''(\theta) = e^{\theta} = \lambda.$$

The canonical link

is $\{g : \eta_i = g(\mu_i) = \theta_i\}$

or $g(b'(\theta_i)) = \theta_i$

So the canonical link
is given by the inverse of

is given by the inverse^{u1}

$$b'(\theta_i)$$

Ex: Normal

$$N(\mu, \sigma^2) : b'(\mu) = \mu \Rightarrow g(\mu) = \mu$$

$$\text{Pois}(\lambda) : b'(\theta) = e^\theta \Rightarrow g(\mu) = \log(\mu)$$

$$\Rightarrow g(\lambda) = \log(\lambda)$$

We don't need to
use the canonical link
process but several nice
...

properties when we do.

GLM Using Canonical Link

Set $\gamma_i = \theta_i = \vec{x}_i^T \beta$

$$f(y_i; \theta_i) = \exp \left\{ \frac{y_i \vec{x}_i^T \beta - b(\vec{x}_i^T \beta)}{a(\phi)} + c \right\}$$

For iid data

$$\mathcal{L}(\beta) = \exp \left\{ \sum \frac{y_i \vec{x}_i^T \beta - b(\vec{x}_i^T \beta)}{a(\phi)} + c \right\}$$

[Aside: $\sum y_i \vec{x}_i$ gives a

S.S. for β

Score function

$$U(\beta) = \frac{d \log L}{d\beta}$$

$$= \frac{1}{a(\phi)} \sum \left(y_i x_i^T - b'(x_i^T \beta) x_i^T \right) = 0$$

$$\Rightarrow \hat{\beta}_{ML} = \left\{ \beta : \sum_i (y_i - \mu_i) x_i^T = 0 \right\}$$

Now to see if this even has a clean solution

We Consider

$$\frac{\partial^2 \log I(\beta)}{\partial \beta \partial \beta^T} = \frac{-1}{\alpha(\phi)} \sum_i x_i b''(x_i^T \beta) x_i^T$$
$$= -\frac{1}{\alpha^2(\phi)} \sum_i \underbrace{x_i V_i x_i^T}_{\text{positive}} \quad \text{(positive definite)}$$

So this thing is

Negative definite for all

β .

$\Rightarrow \mathcal{L}(\beta)$ concave

everywhere

\Rightarrow One unique maximize.

Idea: Solve via iterative

gradient ascent.

Take away: for the right
choice of link the MLE
problem is pretty easy to
solve.

Solve .

Overview

Last Time: GLM mean &
Variance
ML estimation

This Time: ML Estimation
algorithms

Reading: M&N Chp 2

Likelihood Estimation

Putting the canonical
link into the likelihood

$$r_i \leftarrow \sqrt{\lambda} \beta - b(x_i^T \beta) + c$$

$$L(\beta) = \exp \left\{ \frac{1}{a(\phi)} \sum x_i^T \beta - b(x_i^T \beta) + \zeta \right\}$$

$$\Rightarrow U(\beta) = \frac{\partial \log L(\beta)}{\partial \beta}$$

$$= \frac{1}{a(\phi)} \sum y_i (y_i^T - \mu_i^T x_i^T) \stackrel{\text{set}}{=} 0$$

$$\Rightarrow \frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^T} = -\frac{1}{a(\phi)^2} \sum x_i \text{Var}(y_i) x_i^T$$

In matrix notation

$$U(\beta) = \frac{1}{a(\phi)} [y^T - \mu^T] X$$

$$\frac{\partial^2 \log L}{\partial \beta \partial \beta^T} = -\frac{1}{n^2 a(\phi)} X^T V X$$

$$\frac{\partial \mathcal{L}}{\partial \beta \partial \beta^T} = \overline{\alpha^2(\phi)} \quad | \quad \begin{matrix} p \times n & n \times n & n \times p \end{matrix}$$

- This guaranteed us the global \leftarrow local maximum
- Use a gradient ascent algorithm.

Gradient Ascent:

$$\beta^{(k+1)} = \beta^{(k)} + \gamma \nabla \mathcal{L}(\beta^{(k)})^T$$

which

scaling
matrix.

which
direction

Q: What do we choose

for H ? β^0 ?

One popular choice

Fisher
Scoring

$$H^{(k)} = \left(I[\beta^{(k)}] \right)^{-1}$$

$$= - \nabla \left[\frac{\partial \log L}{\partial \beta \partial \beta^T} \right]$$

- Measures the curvature

- of the surface
- The H. s.t. OLS converges in one step.
 - Finding $I(\beta)$ could be difficult. Instead we may wish to use

$$H^{(k)} = I_{obs} (\beta^{(k)})^{-1}$$

$$= - \frac{\partial^L \log L(\beta)}{\partial \beta \partial \beta^T} \Big|_{\beta^{(k)}}$$

Newton's Method

Hence for this alg.

we have

$$\beta^{(k+1)} = \beta^{(k)} + \alpha(\rho) [X^T V^{(k)} X]^{-1} X^T (\gamma - \mu^{(k)})$$

What happens when
we don't use the
Canonical link?

General Link Function

- We require that the link function be invertible

and diff.

• Let

$$l_i(\beta) = \log f_{Y_i}$$

then

$$V_j(\beta) = \sum_i \frac{\partial l_i(\beta)}{\partial \beta_j}$$

$$= \sum_i \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \gamma_i} \cdot \frac{\partial \gamma_i}{\partial \beta_j}$$

$$l_i = \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c$$

$$\text{First: } \frac{\partial l_i}{\partial \theta_i} = \frac{y_i - m_i}{\sim r \sim}$$

$$\overline{\frac{d\theta_i}{d\phi}} = \overline{\frac{1}{\alpha(\phi)}}$$

Second: $\frac{d\theta_i}{d\mu_i} = \left(\frac{d\mu_i}{d\theta_i} \right)^{-1} = \frac{1}{b'(\theta_i)}$

$$= \frac{\alpha(\phi)}{v_i}$$

Third: $\frac{d\mu_i}{d\mu_i} = \left(\frac{d\mu_i}{d\mu_i} \right)^{-1}$

$$= \frac{1}{g''(\mu_i)}$$

Finally: $\frac{d\mu_i}{d\beta_i} = x_{ij}$

Using all this we
can write the score

function as

$$[U(\beta)]_j = \sum_{v_i} \frac{1}{g'(\mu_i)} (y_i - \mu_i)(x_{ij})$$

$$= g'(\mu_i) \sum \underbrace{\frac{1}{v_i g'(\mu_i)^2}}_{\geq 0} (y_i - \mu_i)(x_{ij})$$

$$= \sum_i w_i (y_i - \mu_i)x_{ij} g'(\mu_i)$$

where

$$w_i = \frac{1}{v_i [g'(\mu_i)]^2}$$

and we say $(y_i - \mu_i)$

is error in the data space.

is error in the data space.

Want: Find

$$z_i : (z_i - \eta_i) = (y_i - m_i)g'(m_i)$$

Let

$$z_i = m_i + (y_i - m_i)g'(m_i)$$

$$\Rightarrow V_j(\beta) = \sum_i w_i k_{ij} \underbrace{(z_i - \eta_i)}_{\text{error in linear space.}}$$

error in linear
space.

In matrix notation

then

$$\begin{pmatrix} & \dots \\ & \dots \\ & \dots \end{pmatrix}$$

$$W = \text{diag}(w_1, \dots, w_n)$$

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

$$Z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}, n = X\beta$$

So using this notation

$$U(\beta) = X^T W (Z - X\hat{\beta}) \stackrel{\text{Set.}}{=} 0$$

$$\Rightarrow X^T W Z = X^T W X \hat{\beta}$$

$$\Rightarrow \text{If } (X^T W X)^{-1} \text{ exists}$$

$$\hat{\beta} = (X^T W X)^{-1} X^T W Z$$

• But notice w/z depends
on v_i and $g'(v_i)$

Solution: Iterate until
we find $\hat{\beta}$ that approximately
solves our problem.

Iteratively Reweighted Least Squares

Step 1: Initialize $m^0 = X\beta^0$ or
 $m^0 = g(\gamma)$.

Iterate: k
 $-1, \dots, (k) \rightarrow \text{resp. space}$

Iterate \cdot k

$$-\mu = g^{-1}(\gamma^{(k)}) \quad (\mapsto \text{Resp. Space})$$

$$-W = \text{diag} \left[\left[V(\mu_i) g''(\mu_i)^2 \right]^{-1} \right]$$

$$-Z = \gamma^{(k)} + \text{diag}(g'(\mu))(\gamma - \mu)$$

$$-\beta^{(k+1)} = (X^T W X)^{-1} X^T W Z$$

$$-\gamma^{(k+1)} = X \beta^{(k+1)}$$

- Continue until convergence

$$\left(\gamma^{k+1} - \gamma^k \text{ is small} \right).$$

Overview

Last time: GLM ML
Estimation

This Time: ML Estimation

Inference on β

Goodness of fit

Reading: MN (Chp 3 opt.)

4.1 - 4.4

ML Estimation Ex

$$\hat{\beta} = (\mathbf{x}^T \mathbf{w} \mathbf{x})^{-1} \mathbf{x}^T \mathbf{w} \mathbf{z}$$

$$\beta = (X'WX)^{-1} X'WZ$$

weights
depending
on β

linear
space
Response

(See code: link matters
a lot.)

Inference on β

Let $\hat{\beta}$ be the ML
estimate of β .

Under the Fisher regularity

Conditions

$$\hat{\beta} \xrightarrow{\text{asym.}} N(\beta, I(\beta)^{-1})$$

where

$$I(\beta) = \frac{1}{a(\phi)} (X^T W X)$$

is the Fisher Information

Moreover $\hat{\beta}$ is efficient

in the Fisher sense

This means that for

large n asymptotics

kick in and we can

kick in and we can
use normal type
inference tools (e.g.
hypothesis tests & CI's).

$$-\hat{\beta}_j = \hat{\beta}_j \pm z_{\alpha/2} s_e(\hat{\beta}_j)$$

$$\text{for } s_e(\hat{\beta}_j) = \sqrt{I(\beta)^{-1}}_{jj}$$

To test multiple components
of β simultaneously, use
likelihood ratio tests.

$$\text{E.g. } H_0: \beta = \tilde{\beta}$$

$$H_1: \beta \neq \hat{\beta}$$

$$H_A : \beta = \beta^*$$

$$\text{LogLR} = \frac{1}{a(\phi)} \sum_i \left[y_i (\hat{\theta}_i - \tilde{\theta}_i) - (b(\hat{\theta}_i) - b(\tilde{\theta}_i)) \right]$$

If H_A is nested in H_0 .

Then

$$-2\text{LogLR} \stackrel{\text{asy}}{\sim} \chi^2_{P_0 - P_1}$$

Goodness of Fit

- Need a way to evaluate the model
 - Both deterministic

and stochastic

For SLR we tend to
look at

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

We get this from the
likelihood.

Can we extend this
notion to other Exp
Family dist?

Note: For iid normal

data

$$-2 \log L = \frac{1}{\sigma^2} \sum (y_i - \mu_i)^2 + C$$

Can this help us identify
other GOF statistics?

Want to compare
our model to the
"best" model. E.g.
Saturated model. (n parameter model)

- Compare models via deviance from this model
- . For the proposed model

For the proposed move

$$\hat{m}_i = x_i^T \beta \Rightarrow \hat{m}_i = g^{-1}(\hat{\eta}_i)$$

$$\Rightarrow \hat{\theta}_i = b'^{-1}(\hat{m}_i).$$

For the saturated model

$$\tilde{m}_i = y_i, \hat{\theta}_i = b'^{-1}(\tilde{m}_i)$$

Using these we define

the **model deviance**

$$D(y; \hat{m}_i) := -2 \alpha(\phi) \log LR$$

$$= 2 \sum_i [y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))]$$

Measures how far away

The model in question
is away from saturated
model.

Exercise: Compute

Pcovariance functions

when $Y_i \sim \text{iid Pois}(\lambda)$ and

$Y_i \sim \text{iid } N(\mu, \sigma^2)$

Overview

Last time: Inference on
 β + Goodness of fit

This time: More Goodness of fit and residual analysis

Reading: MC + Neld 4.1 - 4.4

Notes: PS #1 due Wed.

Methods for Goodness of Fit

(1) Deviance:

$$D(y; \hat{\mu}) = -2a(\phi) \log LR$$

$$\boxed{= 2 \sum_i y_i (\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]}$$

where $\tilde{\mu}_i = y_i$ $\tilde{\theta}_i = b'^{-1}(y_i)$

(2) Pearson χ^2 statistic

$$\boxed{\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}}$$

where

$$v(\hat{\mu}_i) = \frac{\text{Var}(y_i | \mu_i)}{a(\phi)}$$

• A sort of weighted sums of

• A sort of weighted sums of squares in the linear space.

Ex: $Y_i \sim \text{iid } N(\mu, \sigma^2)$

y
[]
unknown known

(1) $\theta_i = \mu_i, b(\theta_i) = \frac{\theta_i^2}{2}$

\Rightarrow

$$D(y; \mu) = 2 \sum_i y_i (y_i - \hat{\mu}_i) - \left[\frac{y_i^2}{2} - \frac{\hat{\mu}_i^2}{2} \right]$$

$$= \sum_i y_i^2 - 2y_i \hat{\mu}_i + \hat{\mu}_i^2$$

$$= \sum_i (y_i - \hat{\mu}_i)^2$$

Residual Sum
of Squares.

$$(2) \quad \chi^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

when $V(\hat{\mu}_i) = b''(\hat{\theta}_i) = 1$

So $\chi^2 = RSS$.

Ex: $X_i \sim \text{iid Pois}(\lambda_i) = \text{Pois}(\mu_i)$

$$(1) \quad \theta_i = \log(\mu_i)$$

$$b(\theta_i) = e^{\theta_i}$$

$$p(y_i | \mu) = 2 \sum [y_i [\log y_i - \log \hat{\mu}_i] - [y_i - \mu_i]]$$

$$= 2 \sum_i y_i \log \left(\frac{y_i}{\mu_i} \right) - (y_i - \mu_i)$$

$$-\sum_i y_i \log\left(\frac{y_i}{m_i}\right) - (y_i - m_i)$$

$$(2) V(m_i) = b''(m_i) = m_i$$

$$\chi^2 = \sum_i \frac{(y_i - m_i)^2}{m_i}$$

- This are wholeistic
heuristic gof. What
about individual points?

Residual Analysis

- Many types of definitions
 - Raw Residual

$$r_i = y_i - \hat{m}_i$$

- Working Residual (linear space)

$$r_i = z_i - \hat{\eta}_i$$

$$= (y_i - \hat{m}_i) g'(\hat{m}_i)$$

Most useful:

- Deviance Residuals

$$D_i \approx 2 y_i (\hat{\theta}_i - \hat{\phi}_i) - (b(\hat{\phi}_i) - b(\bar{\phi}))$$

$$r_i = \underbrace{\text{sign}(y_i - \hat{m}_i) \sqrt{D_i}}_{\text{just person sign}}$$

• Note

$$\text{Pch}(\hat{m}) = \sum_i (r_0)_i^2$$

- Pearson Residuals

$$r_i = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}}$$

• Note

$$\chi^2 = \sum_i r_i^2$$

Less Common:

- Anscombe Residuals

$$\text{Let } A(\mu) = \int_0^\mu V^{-\frac{1}{2}}(\tilde{\mu}) d\tilde{\mu}$$

then

$$r_i = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i) \sqrt{V(\hat{\mu}_i)}}$$

Ex: $y_i \sim N(\mu_i; \sigma^2)$

Raw

$$r_i = y_i - \hat{\mu}_i$$

Raw

$$r_i = y_i - \mu_i$$

Working

$$r_i = y_i - \hat{\mu}_i$$

Deviance

$$r_i = y_i - \hat{\mu}_i$$

Pearson

$$r_i = y_i - \hat{\mu}_i$$

Ascombe

$$\beta(\mu) = \mu \Rightarrow$$

$$r_i = y_i - \hat{\mu}_i$$

Ex: $y_i \sim \text{iid Pois}(\mu_i)$

Raw

$$r_i = y_i - \hat{\mu}_i$$

Working

$$r_i = \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}$$

Deviance

$$r_c = \text{sign}(y_i - \hat{\mu}_i) \left[2 y_i \log \frac{y_i}{\hat{\mu}_i} - [y_i - \hat{\mu}_i] \right]^{1/2}$$

Pearson

$$r_i = \frac{(y_i - \hat{m}_i)}{\sqrt{\hat{n}_i}}$$

Ascombe

$$A(n) = \int_0^n \hat{m}^3 d\hat{m} = \frac{3}{2} n^{2/3}$$

$$r_i = \frac{3}{2} \left[\frac{y_i^{2/3} - \hat{m}_i^{2/3}}{\hat{m}_i^{1/3} \hat{M}_i^{1/2}} \right]$$

$$= \frac{3}{2} \left[\frac{y_i^{2/3}}{\hat{m}_i^{2/3}} - \hat{M}_i^{1/2} \right]$$

Key note: Which residuals

do we use in each situation?

How do they help us identify

(lack of) model fit performance?

What to do with residuals?

- (1) Visualize them
- (2) Q-Q plots
- (3) Autocorrelation structure.
- (4) Cross correlation between residuals & covariates

Lecture 2/7

Wednesday, February 7, 2018 1:24 PM

Overview

Last Time: Residual Analysis

This Time: GLM Theory for
Binary Data

Reading: M&N chp 4

Faraway 2

Modeling Binary Response

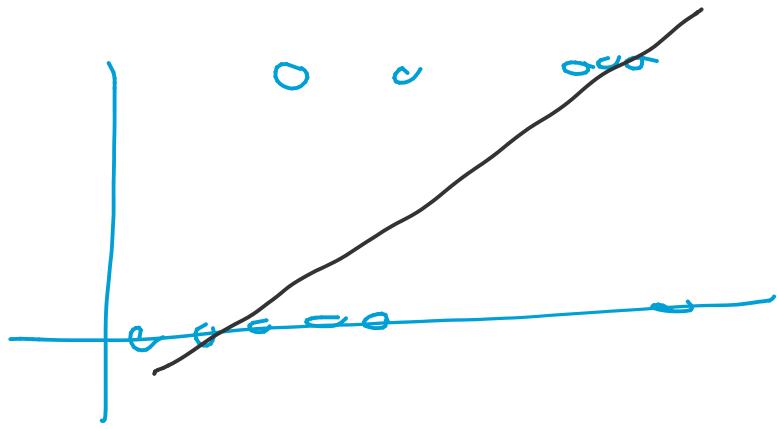
- Nonbinary data can be dichotomized

$$Y < \lambda \text{ vs. } Y \geq \lambda$$

Good for modeling data

We aren't entirely sure
falls into an exponential
family.

- SLR bad here b/c



- Bad CI, inference, non-helpful coefficient.
- 2 dists. of interest
 Bernoulli or Binomial

Bernoulli Model -

$Y_i \sim_{iid} \text{Bern}(p_i)$ where

p_i depends on x_i

$$P(Y_i = y_i) = p_i^{y_i} (1-p_i)^{1-y_i}$$

$$= \exp \left\{ y_i \log(p_i) + (1-y_i) \log(1-p_i) \right\}$$

$$= \exp \left\{ y_i [\log(p_i) - \log(1-p_i)] + \log(1-p_i) \right\}$$

$$= \exp \left\{ y_i \log \left(\frac{p_i}{1-p_i} \right) + \log(1-p_i) \right\}$$

Define $\theta_i = \log \left(\frac{p_i}{1-p_i} \right)$

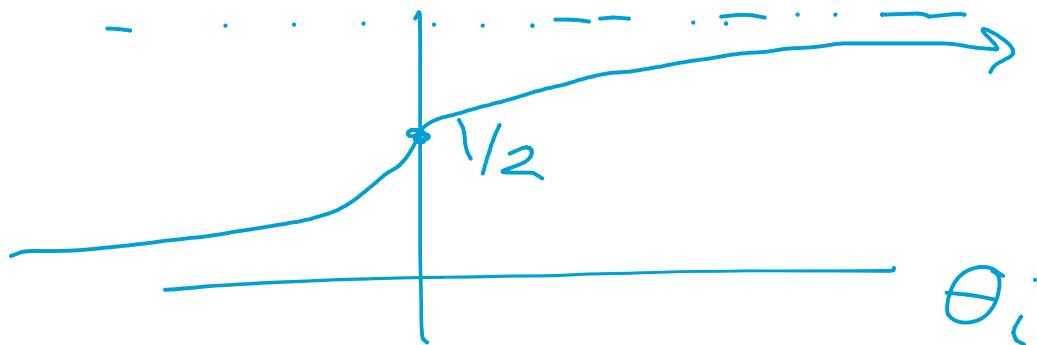
$$:= \text{logit}(p_i)$$

$$\Rightarrow p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$$

sigmoid

$$\Rightarrow p_i = \frac{e^{\theta_i}}{1 + e^{\theta_i}}$$

Sigmoid
function



Confidence bounds on θ_i will hence give C.I. between $(0, 1)$

for p_i .

Now we also have

$$g(\phi) = 1$$

$$b(\theta) = -\log(1-p) =$$

$$= \log\left(\frac{1}{1-p}\right)$$

$$= \log(1+e^\theta)$$

$$\mu_i = b'(\theta) = \frac{e^{\theta_i}}{1+e^{\theta_i}} = p_i$$

\Rightarrow Canonical link

$$g(\mu_i) = \theta_i = \log \frac{p_i}{1-p_i}$$

$$= \log\left(\frac{\mu_i}{1-\mu_i}\right) = \text{logit}(\mu_i)$$

θ

- 26

$$b''(\theta_i) = \frac{e^\theta}{1+e^\theta} - \frac{e^{2\theta}}{(1+e^{\theta_i})^2}$$

$$= p_i - p_i^2 = p_i(1-p_i) = \text{Var}(Y_i)$$

Binomial Model

$$Y_i \sim \text{iid} \text{ Binom}(p_i, n_i)$$

↑ ↑
unknown. known

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i-y_i}$$

$$= \exp \left\{ y_i \log \left(\frac{p_i}{1-p_i} \right) + n_i \log (1-p_i) \right. \\ \left. + \dots / n_i \right\}$$

$$+ \log \left(\frac{n_i}{y_i} \right) \}$$

$$\theta_i = \text{logit}(p_i)$$

$$a(\chi) = 1, \quad b(\theta_i) = n_i \log(1 + e^{\theta_i})$$

$$b'(\theta_i) = n_i p_i$$

\Rightarrow Canonical link

$$g(\mu_i) = \theta_i = \log \frac{p_i}{1-p_i}$$

$$= \log \left(\frac{m_i/n_i}{1-m_i/n_i} \right) = \log \left(\frac{m_i}{n_i - m_i} \right)$$

$$b''(\theta_i) = n_i p_i(1-p_i)$$

$$= m_i \frac{(n_i - m_i)}{n_i} = V_i$$

Building a Binary GLM.

Using the canonical link, set

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta}$$

\Rightarrow

$$p_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

$$P_i = \frac{e^{x^T \beta}}{1 + e^{x^T \beta}}$$

Fit by either Fisher or
Newton Scoring or I.R.L.S.

To give $\hat{\beta}_{ML}$.

Let $\frac{P_i}{1-P_i} \leftarrow \text{Success} = \text{odds}_i = o_i$

Interpret as a betting
ratio. So the natural
parameter is just the

parametric - \cup

log-odds.

Ex: (Binary Predictors)

y_i = disapproval of president

$$x_i = \begin{cases} 0 & \text{Democrat} \\ 1 & \text{Republican} \end{cases}$$

$$p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Democrat Model

$$p_{i,D} = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \quad \text{so}$$

β_0 is log-odds that a democrat disapproves of the president.

Republican Model

$$P_j = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \Rightarrow o_j = e^{\beta_0 + \beta_1}$$

$$\text{So } \frac{o_j}{o_i} = e^{\beta_1} \Rightarrow$$

log (odds ratio) modulation
 in odds between republicans

and democrats.

If $\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \log 2 \\ \log 1/4 \end{pmatrix}$

$$= \begin{pmatrix} 0.69 \\ -1.39 \end{pmatrix}$$

\Rightarrow odds that a democrat

$$2:1 \Rightarrow P_{\text{dem}} = 2/3$$

$$\text{odds ratio} = 1/4$$

$$\text{Rep odds} = \frac{2:1}{4} = \frac{1}{2} = P_{\text{rep}} = 1/3$$

Ex: (Single Cont. Pred)

Ex. (Single variable)

$$Y_i = \begin{cases} 0 & \text{no frost bite} \\ 1 & \text{frostbite} \end{cases}$$

X_i = temp in $^{\circ}\text{C}$

$$P_i = \frac{\exp\{\beta_0 + \beta_1 X_i\}}{1 + \exp\{\beta_0 + \beta_1 X_i\}}$$

$\Rightarrow \beta_0$ logodds of frostbite
if $X_i = 0$

e^{β_1} = odds ratio when

X_i increases by 1°

odds of frost bite when

$X_i = -30$ is given by

$$\exp\left\{\beta_0 + \beta_1(-30)\right\}$$

$$= (e^{\beta_0})(e^{\beta_1})^{-30}$$

Lecture 2/9

Friday, February 9, 2018 1:24 PM

Overview

Last Time: Binomial GLM Theory

This Time: Fitting logistic Reg.

Inference on log.

Reef.

Fitting Logistic Reg.

If $Y_i \sim \text{Binom}(n, p_i)$

Set

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta$$

Logistic Regression model
 \iff

Binomial GLM with canonical link.

Ex: Frostbite on Everest

x_i = temp at peak

$$y_i = \begin{cases} 0 & \text{no frostbite} \\ 1 & \text{frostbite} \end{cases}$$

$$P_i = \frac{\exp\{x_i^T \beta\}}{1 + \exp\{x_i^T \beta\}}$$

e^{β_0} = odds of frost bite
at $x_i = 0$

$e^{\beta_1} = \frac{O_{X+1}}{O_X} =$ change in
odds for
each 1 degree
 C° increase.

Suppose we know

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \ln(Y_{100}) \\ 0 / -1 \end{pmatrix} = \begin{pmatrix} -4.61 \\ -0.105 \end{pmatrix}$$

$$(\beta_1) = (\ln(0.9)) \quad (-0.105)$$

$$\Rightarrow P_{0^\circ C} = \frac{1}{100}$$

Or the prob someone gets frost bite at $0^\circ C$ is

$$\frac{1}{100+1} = \frac{1}{101}$$

Similarly

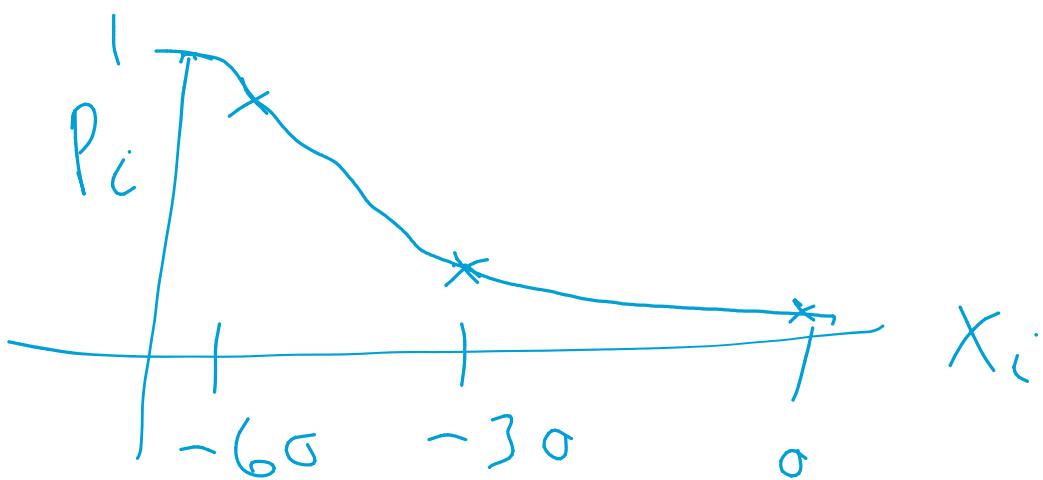
$$P_{-30^\circ C} = \frac{1}{100} (0.9)^{-30} = 0.24$$

$$\Rightarrow P_{-30^\circ C} = \dots 19$$

$$O_{-60^\circ} = \frac{1}{100} (0.9)^{-60} = 5.6.$$

$$P_{-60^\circ} = \frac{5.6}{1+5.6} = 0.85$$

Plotting this behavior



Estimation IRLS

$$V_i = n_i p_i (1-p_i)$$

$$= \frac{m(n-m)}{n}$$

$$g(m_i) = \log \frac{m}{n-m}$$

$$g'(m_i) = \frac{n-m}{m} \cdot \left(\frac{1}{n_i - m_i} + \frac{m_i}{(n_i - m_i)^2} \right)$$

$$= \frac{n_i - m_i}{m_i} \left(\frac{n_i - m_i + m_i}{(n_i - m_i)^2} \right)$$

$$= \frac{n_i}{m_i(n_i - m_i)} = \frac{1}{V_i}$$

Now looking at the

Now looking at the weight matrix

$$W_i = \frac{1}{V_i [g'(m_i)]^2} = V_i$$

IRLS algorithm.

Initialize $P_i^{(0)} = \frac{x_i + y_i}{n_i + 1}$

to prevent initial values
of 0/1.

Then get linear space
initial estimate

$$\gamma_i^{(0)} = \log \left(\frac{p_i^{(0)}}{1-p_i^{(0)}} \right)$$

Iterate:

$$- p_i^{(k)} = \frac{e^{\gamma_i^{(k)}}}{1 + e^{\gamma_i^{(k)}}}$$

$$- \boldsymbol{\mu} = \begin{pmatrix} n_1 p_1^{(k)} \\ \vdots \\ n_r p_r^{(k)} \end{pmatrix}$$

$$- \boldsymbol{W} = \text{diag}(n_i p_i^{(k)} (1 - p_i^{(k)}))$$

$$- \boldsymbol{\beta} = \boldsymbol{\gamma}^{(k)} + (\boldsymbol{y} - \boldsymbol{\mu}) \boldsymbol{V}^{-1}$$

$$- \boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\gamma}$$

$$-\gamma^{(k+1)} = X\beta^{(k+1)}$$

Notice here that

$$W = V \text{ so } (X^T W X)^{-1}$$

is Fisher information

so

$$\hat{\beta} \stackrel{\text{asy}}{\sim} N(\beta, (\bar{X}^T W X)^{-1})$$

Ex: $X_i = \begin{cases} 0 & \text{democrat} \\ 1 & \text{republican} \end{cases}$

$$n_i = \begin{cases} 100 \\ 50 \end{cases} \quad Y_i = \text{Pres. disapprov.}$$

$$= \begin{cases} 70 \\ 15 \end{cases}$$

$$\Rightarrow \hat{\beta} = \begin{pmatrix} 0.85 \\ -1.7 \end{pmatrix} \quad \begin{matrix} \rightarrow \text{prob that} \\ \text{dem. disapprov} \\ > Y_2 \end{matrix}$$

adds a

republican

disapproves

Mercusg.

$$I(\hat{\beta})^{-1} = \begin{bmatrix} 0.05 & -0.05 \\ -0.05 & 0.14 \end{bmatrix}$$

$$\text{“(P)"} \quad \left[-0.05 \quad 0.14 \right]$$

$$\hat{O}_D = e^{0.85} = \frac{2.3}{1}$$

$$\hat{P}_D = \frac{2.3}{1+2.3} = 0.7$$

$$\hat{O}_R \text{ R.V.D.} = e^{-1.7} = 0.18$$

$$\hat{O}_R = (2.3)(0.18) = \frac{0.43}{1}$$

$$\hat{P}_R = \frac{0.43}{1+0.43} = 0.3$$

$$\text{Var}(\hat{\beta}_0) = 0.05$$

\Rightarrow A 95% CI can
be given as.

$$\hat{\beta}_0 \pm z_{\alpha/2} \sigma_{\hat{\beta}_0}$$

$$(0.42, 1.27)$$

So 95% Confidence that

$$P_0 > 1/2$$

$$\text{Var}(\hat{\beta}_1) = 0.14$$

$$\text{Var}(\beta_1) = 0.14$$

$$\Rightarrow -1.7 \pm z_{\alpha/2} \sqrt{0.14}$$

$$(-2.4, -0.95)$$

\Rightarrow We have evidence

that $OR < 1 \Rightarrow P_D \neq P_R$.

95% CI on P_D

Let the CI of β_0

be (l_0, r_0) . Then the

CI is given by

Let us go on

$$\left(\frac{e^{l_o}}{1+e^{l_o}} + \frac{e^{r_o}}{1+e^{r_o}} \right)$$

For $P_R = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$

So we need CI on

$$\beta_0 + \beta_1$$

$$\text{Var}(\beta_0 + \beta_1) = 0.05 + 0.14 - 2(0.05) = 0.09$$

So this CF is

$$(0.85 - 1.7) \pm 2\% \sqrt{0.09}$$

$$\therefore = (\ell_1, r_1)$$

$$P_R = \left(\frac{e^{\ell_1}}{1 + e^{\ell_1}}, \frac{e^{r_1}}{1 + e^{r_1}} \right)$$

$$= (0.19, 0.44)$$

Overview

Last Time: Fitting/

inference for Logistic

This Time: Choice of links

Prospective vs. Retrospective

Sampling

Notes: No class Friday

Choice of Links

The Logit is the
Canonical link for

Binomial

Other feasible choices.

Four most common.

1. Logit $g(p_i) = \log \frac{p_i}{1-p_i}$

$$p_i = \frac{\exp(m_i)}{1 + \exp(m_i)}$$

2. Probit $g(p_i) = \Phi^{-1}(p_i)$

< 1. (Logit) $g(\eta_i) = \Phi(p_i)$

$$\underline{\Phi}(x) = P(Z \leq x)$$

$$p_i = \underline{\Phi}(\eta_i)$$

3. Complementary

$$\text{Log-Log } g_c(p_i) = \ln(-\log(1-p_i))$$

$$p_i = -\exp(-e^{\eta_i})$$

4. Log-log $g(p_i) = -\log(-\log(p_i))$

$$p_i = \exp\{-e^{-\eta_i}\}$$

why would we choose

... "J vu- - -"

one over the canonical
link?

- Tail behaviors

- Logit has long tails
 - slow-left tail, fast-right tail
 - fast-left tail

left tail slow
right tail.

- Link usually chosen for its properties in each model.

Pro/Retro prospective Sampling

Ex: Imagine a 2×2 Contingency table

D.	D
----	---

	D	D
X	n_{11}	n_{12}
X	n_{21}	n_{22}

"—" = does not happen.

Prospective: do you smoke

\Rightarrow
did you get
the disease

Retrospective: You do (n't)
have the disease

\Rightarrow
did you smoke?

did you smoke?

For a logit regression

$$D \sim X_i \quad \hat{\beta}_i = \log \frac{\hat{P}_{DX}}{\hat{P}_{DX}}$$

$$= \log \frac{n_{22}/n_{-2}}{n_{12}/n_{-1}}$$

$$= \log \frac{n_{22}n_{-1}}{n_{21}n_{12}}$$

For a model

$$X \sim D \quad \hat{\beta}_i = \log \frac{\hat{P}_{XID}}{\hat{P}_{XID}}$$

$$= \log \frac{n_{22}/n_{12}}{n_{21}/n_{-1}}$$

$$\sigma_{n_2/n_1} = \log \frac{n_{21} n_{11}}{n_{21} n_{12}}$$

So $\hat{\beta}_1$ in both models

are the same. That is

inference on the association
will be the same.

- Note this is only true for choice of the canonical link

function.

More generally:

Prospective model

$$P(R_i | \vec{x}_i) = \frac{\exp(\alpha + \vec{x}^T \beta)}{1 + \exp(\alpha + \vec{x}^T \beta)}$$

Retrospective model

$$\text{Let } z_i = \begin{cases} 1 & \text{sampled} \\ 0 & \text{not sampled} \end{cases}$$

Assume $P(z_i = 1 | R_i, x_i)$

$$= P(z_i = 1 | R_i) = \begin{cases} \pi_p & R_i = 1 \\ 1 - \pi_p & R_i = 0 \end{cases}$$

$$= P(z_i = 1 | D_i) = \begin{cases} \pi_D & p_i = 1 \\ \pi_{\bar{D}} & p_i = 0 \end{cases}$$

$$P(D_i | X_i, z_i = 1)$$

$$= \frac{P(z_i = 1 | D_i, X_i) P(D_i | X_i)}{P(z_i | D_i, X_i) P(D_i | X_i) + P(z_i | \bar{D}_i, X_i) P(\bar{D}_i | X_i)}$$

$$= \frac{\pi_D \frac{\exp(\alpha + X_i^T \beta)}{1 + \exp(\alpha + X_i^T \beta)}}{\pi_D \frac{\exp(\alpha + X_i^T \beta)}{1 + \exp(\alpha + X_i^T \beta)} + \pi_{\bar{D}} \frac{1}{1 + \exp(\alpha + X_i^T \beta)}}$$

$$= \frac{\pi_D \exp(\alpha + X_i^T \beta)}{\pi_D \exp(\alpha + X_i^T \beta) + \pi_{\bar{D}} \frac{1}{1 + \exp(\alpha + X_i^T \beta)}}$$

$$= \frac{\pi_D \exp(\alpha + X_i^T \beta)}{\pi_D \exp(\alpha + X_i^T \beta) + \pi_{\bar{D}}}$$

$$\pi_p \exp(\alpha + x_i^\top \beta) + \bar{\pi}_{\bar{D}}$$

$$= \frac{\pi_D / \bar{\pi}_{\bar{D}}}{1 + \frac{\pi_D}{\bar{\pi}_{\bar{D}}} \exp(\alpha + x_i^\top \beta)}$$

$$= \frac{\exp(\alpha^* + x_i^\top \beta)}{1 + \exp(\alpha^* + x_i^\top \beta)}$$

$$= \frac{\exp(\alpha^* + x_i^\top \beta)}{1 + \exp(\alpha^* + x_i^\top \beta)}$$

where $\alpha^* = \alpha + \log \frac{\pi_p}{\bar{\pi}_{\bar{D}}}$

Notice • β 's are unchanged

- Δ^* intercept changes.
- Great for Sampling schemes.

Goodness of Fit

Deviance:

$$D(y; \hat{\mu}) = 2 \log L(\hat{\mu}; y) - 2 \log L(\bar{\mu}; y)$$

$$= 2 \sum_i y_i \log \left(\frac{y_i}{n_i} \right) + (n - y_i) \log \left(1 - \frac{y_i}{n_i} \right)$$

$$-(n_i - y_i) \log\left(1 - \frac{\hat{m}_i}{n_i}\right)$$

$$= 2 \sum_i y_i \log\left(\frac{y_i}{\hat{m}_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{m}_i}\right)$$

success weight.
fail weight
successes
failures

This gives the residual

$$(r_D)_i =$$

$$\text{sign}(y_i - \hat{m}_i) \left[y_i \log\left(\frac{y_i}{\hat{m}_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{m}_i}\right) \right]^{1/2}$$

Pearson's χ^2 gives

Fearson's χ^2 gives

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V_i} = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\frac{\hat{\mu}_i(n_i - \hat{\mu}_i)}{n_i}}$$

$$= \sum_i n_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)}$$

\Rightarrow

$$(F_X)_i = \left(\frac{n_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} \right)^{\frac{1}{2}} (y_i - \hat{\mu}_i)^2$$

Lecture 2/14

Wednesday, February 14, 2018 1:25 PM

Overview

Last time:

- Choice of Link
- Pro/Retro perspective

This time:

- GOF
- Stepwise Reg.

Reading MAN Chp 5

Reading MAN Lhp J

Goodness of Fit

$X_i \sim \text{ind Binom}(n_i, p_i)$

$$g(p_i) = x_i^T \beta$$

$$D(\hat{y}_i; \hat{\pi}) = \sum_i \left(y_i \log \left(\frac{y_i}{\hat{n}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{n}_i} \right) \right)$$

Absolute GOF

Hypothesis testing

If the total number

of groups is fixed

$$\dim(Y) = N$$

As $n_i \rightarrow \infty$

$$p(y_{ij\hat{n}}) \xrightarrow{\text{asy}} \chi^2_{n-p}$$

Useful an absolute value of fit.

Relative GOF

Comparing nested logistic
Regression Models.

$$H_0: X_0 = \begin{bmatrix} x_{01} & \dots & x_{0p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$H_A: X_A = \begin{bmatrix} X_0 | \vec{X}_n \end{bmatrix} \quad \gamma = \begin{bmatrix} r_1 \\ \vdots \\ r_p \\ \gamma_A \end{bmatrix}$$

Assume H_0 is the correct model i.e. $\gamma_A = 0$.

Which model has smaller deviance?

H_A . By setting $\hat{\gamma}_i = \hat{\beta}_i$

& $\tilde{\gamma}_A = 0$ the models

attain the same deviance.

attain the same deviance.

But $\hat{\beta}$ may not even be
maximum likelihood.

If H_0 true

$$\Delta D = D(y; \hat{\beta}_{mL}) - D(y; \hat{\beta}_{ML})$$

$\sim \chi^2_1$

More generally,

$$H_0: \{\beta : \beta_{e_1} = \dots = \beta_{e_h} = 0\}$$
$$H_A: \{ \exists k. \text{ s.t. } \beta_k \neq 0 \}$$

ΔD ^{asym.} χ^2_k

Stepwise Regression

Modeling construction

VRN S.R. Let's suppose

we have a collection

of covariates $\{X^{(k)}\}$

Start with null model

$$X^0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \eta^0 = X^0 \beta^0$$

Fit $\hat{\beta}_0$ and compute deviance.

Then one-by-one include new covariate

$\{X^{(k)}\}$, calculate deviance, and test $H_0: \Delta D = 0$

~ 2

UV-- · --

$H_A: \Delta D \neq 0$ via $\chi^2_{1,2}$

Can be rapidly visualized
with A \circ D Deviance (A \circ D) table

Model	Tot. df	Δdf	Dev	ΔDev	P
Null	$N - 1$	0	-	-	-
χ^2	$N - 2$	1	.	.	.
.
.
$[x_1 \dots x_p]$	$N - P$	$P - 1$.	.	.

Note: Order matters

unless independent.

Akaike Information Criterion

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

↑
likelihood ↑
Penalty

Remark: Minimizing

AIC \Leftarrow Complete C.V.

Overview

Last time: GoF for binary
GLM

This time: Overdispersion

Reading: M&N Chp. 5

Overdispersion

Let $Y_i \sim \text{Binom}(n_i, p_i)$

If p_i are known then

$V_i = n_i p_i (1-p_i)$. So we can

$V_i = n_i p_i(1-p_i)$. So we can

extend our estimate of
 p_i to get estimates of the
Variance \hat{V}_i

For example, we expect.

$$(r_p)_i = \frac{y_i - \hat{m}_i}{\sqrt{\hat{V}_i}} \text{ to}$$

have approximately unit
variance.

In practice we normally
see that $\hat{V}_i \gg 1$.

See that $\text{Var} \sim \perp$.

Suggests $\text{Var}(Y_i) > \hat{V}_i$

which we call overdispersion

Less often $\text{Var}(Y_i) < \hat{V}_i$

which we call underdispersion

How can we combat
overdispersion?

What might explain
overdispersion?

1. Unmodeled variability

in ρ_i

2. Correlated responses X_i .

What will actual variance look like in each case

2) Let $X_i = \sum_{j=1}^{n_i} R_{ij}$

$$R_{ij} \sim \text{Bern}(\rho_i)$$

$$\text{Corr}(R_{ij}, R_{ik}) = \alpha$$

$$\text{Cov}(R_{ij}, R_{ik}) = \alpha \rho_i(1 - \rho_i)$$

This then gives

$$\text{Var}(Y_i) = \sum_{j=1}^{n_i} \text{Var}(R_{ij})$$

$$+ \sum_{j=1}^{n_i} \sum_{k \neq j} \text{Cov}(R_{ij}, R_{ik})$$

$$= n_i p_i (1-p_i) + n_i (n_i - 1) \alpha p_i (1-p_i)$$

$$= n_i p_i (1-p_i) \left[1 + (n_i - 1) \alpha \right]$$

extra
variability

1) Let p_i be random

$$\text{Var}(Y_i) = \mathbb{E}[\text{Var}(Y_i | p_i)]$$

$$+ \text{Var}(\mathbb{E}(Y_i | p_i))$$

$$= \mathbb{E}[n_i p_i (1-p_i)]$$

$$+ \text{Var}(n_i p_i)$$

$$= n_i [\mathbb{E}(p_i) - \mathbb{E}(p_i^2)] + n_i^2 \text{Var}(p_i)$$

$$= n_i [\mathbb{E}(p_i) - \text{Var}(p_i) + \mathbb{E}(p_i)^2]$$

$$+ n_i^2 \text{Var}(p_i)$$

$$= h_i [\mathbb{E}(p_i) - \mathbb{E}(p_i)^2] \\ + n_i(n_i - 1) \text{Var}(p_i)$$

Assume that

$$\text{Var}(p_i) = \alpha \mathbb{E}(p_i)[1 - \mathbb{E}(p_i)]$$

This gives

$$\text{Var}(Y_i) = n_i \mathbb{E}(p_i)(1 - \mathbb{E}(p_i))$$

$$+ n_i(n_i - 1) \alpha \mathbb{E}(p_i)(1 - \mathbb{E}(p_i))$$

$$= n_i \mathbb{E}(p_i)(1 - \mathbb{E}(p_i)) \left\{ 1 + \underbrace{(n_i - 1)\alpha}_{\text{Overdispersion}} \right\}$$

\hat{T}

Tests for Overdispersion:

Assume

$$\text{Var}(Y_i) = \sigma^2 n_i p_i (1-p_i)$$

$$H_0: \sigma^2 = 1 \quad H_A: \sigma^2 > 1$$

Let

$$\hat{\sigma}^2 = \frac{\chi^2}{N-p}$$

then

$$(N-p) \frac{\hat{\sigma}^2}{\sigma^2} \xrightarrow{\text{asy}} \chi^2_{N-p}$$

Under the null...

$$(N-p) \frac{\hat{\sigma}^2}{\sigma^2} = \chi^2 \text{ t.s.}$$

which leads to

Significant overdispersion

if $\chi^2 > \chi^2_{N-p, 1-\alpha}$.

Note: Estimates of β are unaffected by under/overdisp.

But it does affect the Fisher information

$$I(\mu) = \frac{1}{\sigma^2} (X^T W X)$$

Inference using $\hat{\sigma}^2$ is
- II I binomial b/c

called quasi-binomial b/c

Covariance.

Tests on individual parameters

using

$$se_{\hat{\beta}_i} = \sqrt{(\hat{\sigma}^2 (X^T W X)^{-1})_{ii}}$$

$$t = \frac{\hat{\beta}_i - \beta^*}{se_{\hat{\beta}_i}} \sim t_{N-p}$$

$$se_{\hat{\beta}_i}$$

Analysis of Deviance for
nested models. we use

$$F = \frac{D(y_i; \mu_i^{(z)}) - P(y_i > \hat{\mu}_i^{(z)})}{P - Q}$$

$\hat{\sigma}^2$
from model (z)

$$\sim \frac{\chi^2_{P-Q}}{\chi^2_{N-P}} \sim F_{P-Q, N-P}$$

Overview

Last Time: Binomial Overdisp.

This Time: Multinomial

Reading: M&N 6.1-6.3

Multinomial

Let M_i be a discrete

R.V. over $\{1, 2, \dots, m\}$

with $P(M_i=j) = P_{ij}$

For the i^{th} experiment

we have n_i repetitions

and y_{ij} fall into category

j.

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{im} \end{pmatrix} \sim \text{Multi}\left[n_i, \begin{pmatrix} p_{i1} \\ \vdots \\ p_{im} \end{pmatrix}\right]$$

with

$$P[y_{i1}=y_{i1}, \dots, y_{im}=y_{im}]$$

$$= \frac{n_i!}{y_{i1}! \dots y_{im}!} p_{i1}^{y_{i1}} \dots p_{im}^{y_{im}}$$

$$y_{i1}! \dots y_{im}!$$

$$\text{where } \sum_{k=1}^m p_{ik} = 1$$

$$\sum_{k=1}^m Y_{ik} = n_i$$

Ex: Put into canonical form.

Categories

- Nominal (no ordering)
e.g. Blood Type
- Ordinal (ordering)
e.g. movie ratings
- Hierarchy
e.g. Tree structured
 1 + m

e.g.

data

For nominal categories

(i.e. any multinomial data)

Define a multinomial

logit model:

$$P_{ij} = \frac{\exp\{\gamma_{ij}\}}{\sum_{k=1}^m \exp\{\gamma_{ik}\}}$$

Let $\gamma_{i1} = 0$ as group

1 will serve as a baseline.

$$P_{i1} = \frac{\exp\{0\}}{\sum_{k=1}^m \exp\{\gamma_{ik}\}} = \frac{1}{\sum_{k=1}^m \exp\{\gamma_{ik}\}}$$

$$K_{i,:} = \frac{\sum_{k=1}^m \exp\{\gamma_{ik}\}}{\sum_{k=1}^m \exp\{\gamma_{ik}\}}$$

From here

$$P_{ij} = e^{\gamma_{ij}} \cdot P_{i,1}$$



$$\gamma_{ij} = \log\left(\frac{P_{ij}}{P_{i,1}}\right)$$

log odds of grp j

v.s. gp 1.

Ex: Let $j = \begin{cases} 1 & \text{Dem} \\ 2 & \text{Rep} \\ 3 & \text{Ind.} \end{cases}$

Let $X_i = \text{Annual income}$

Let X_i = Annual income
in \$1000

$$\gamma_{ij} = \theta_j + \beta_j X_i$$

log odds
of j vs.
dem. at
 $X_i = 0$

For each
extra 1000
how that changes
in log odds of
 j vs. 1.

If we obtain fit
parameter values.

$$\hat{\theta}_2 = -0.9 \quad \hat{\theta}_3 = -1$$

$$\hat{\beta}_2 = 0.02 \quad \hat{\beta}_3 = 0.01$$

At $x_i = 0$.

$$P_{D,0} = \frac{1}{1 + e^{-0.9} + e^{-1}} \approx 0.56$$

$$P_{R,0} = P_{D,0} e^{-0.9} \approx 0.23$$

$$P_{I,0} = P_{D,0} e^{-1} \approx 0.21$$

At $x_i = 100$

$$P_{D,100} = \frac{1}{1 + e^{-0.9} + e^{-1}} \approx 0.2$$

$$P_{R,100} = P_{D,100} e^{-0.9(0.2)(100)} \approx 0.6$$

$$P_{I, \text{wo}} = P_{P, \text{wo}} e^{-1} e^{(0.41)(100)} \approx 0.2$$

Parallel Regression for

Ordinal Data:

Instead of modeling

P_{ij} we instead model

$$\gamma_{ij} = P(m_i \leq j)$$

$$\text{Let } \gamma_{ij} = \frac{e^{m_{ij}}}{1 + e^{m_{ij}}} \text{ for } j=1, \dots, m-1$$

$$\gamma_{im} = P(m_i \leq m) = 1 .$$



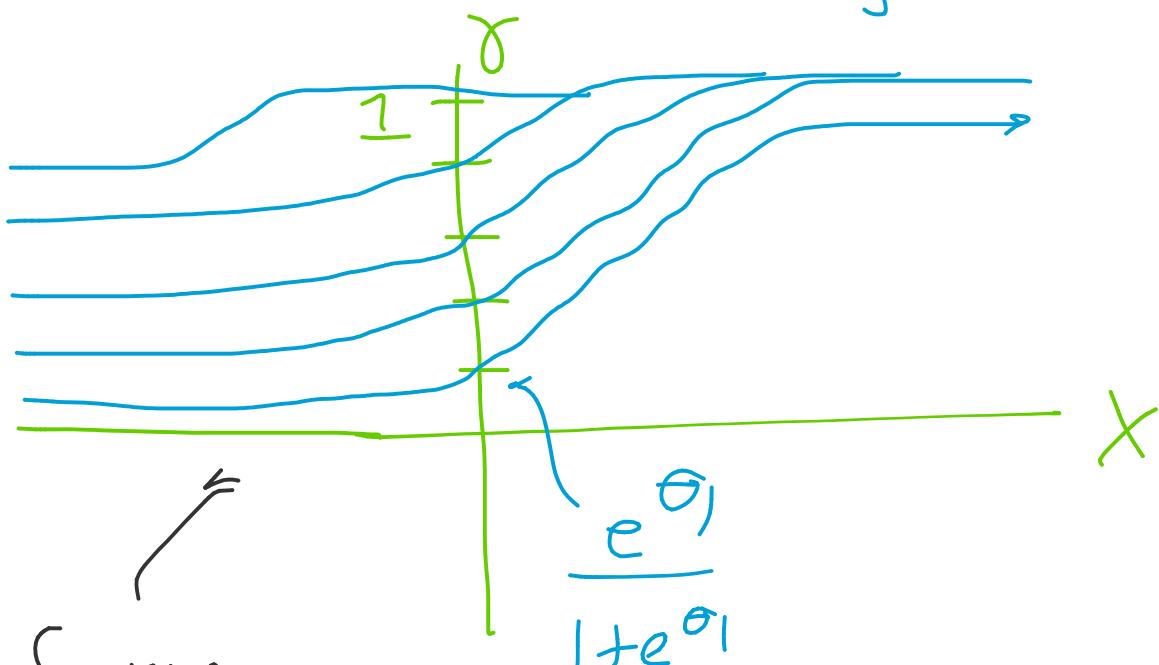
$$m_{ij} = \log \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right)$$

Common model form:

$$\eta_{ij} = \theta_j + X_i^T \beta$$

\uparrow independent
of j

$$(\theta_1 \leq \theta_2 \leq \dots \leq \theta_{n-1}).$$



Same Sigmoid
- Just Shifted.

Proportional
Odds model.

Shifted.

Ex: Use ordering

Penn < Ind < Rep

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \theta_j + \beta x_i$$

$$\hat{\theta}_1 = 0.2 \quad \hat{\theta}_2 = 1.2$$

$$\hat{\beta} = -0.016$$

At income $x_i = 0$

$$P_{D,\sigma} = \frac{e^{0.2}}{1 + e^{0.2}} \approx 0.55$$

$$P_{I,\sigma} = \frac{e^{1.2}}{17} - P_{D,\sigma} = 0.22$$

$$P_{I,\sigma} = \frac{e}{1+e^{1.2}} - P_{R,\sigma} = 0.22$$

$$P_{R,\sigma} = 1 - 0.55 - 0.22 = 0.23$$

At income $x_i = 100$

$$P_{D,100} = \frac{e^{0.2 - 0.016}}{1 + e^{0.2 - 0.016}} = 0.2$$

$$P_{I,100} = 0.2 \quad P_{R,100} = 0.6$$

Almost the same probabilities
except one fewer parameter.

Overview

Last time: Multinomial
Data

This time: Poisson Data

Reading: M&N Chp. 6

Poisson Probability Model

$$Y \sim \text{Pois}(\lambda)$$

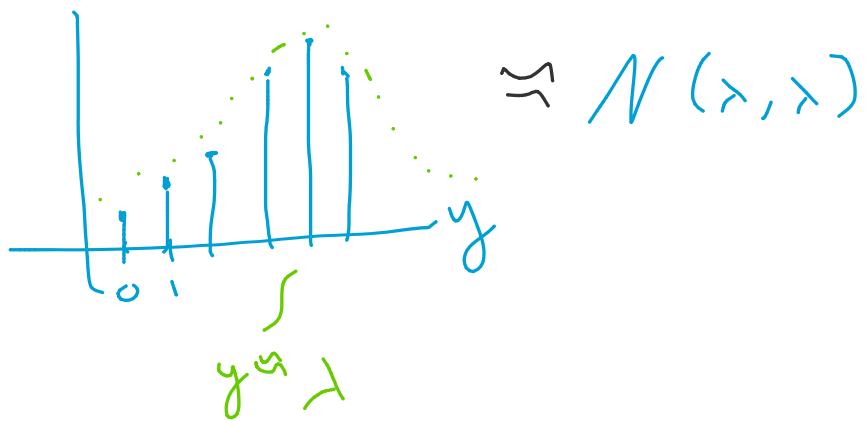
$$P(Y=y) = \frac{\lambda^y e^{-\lambda}}{y!}, y=0, 1, 2, \dots$$

λ small



~~Hilberty~~ with small p

λ large



Used primarily to count independent events that occur continuously in either time or space.

We'll think about Poisson processes primarily.

Partition $[0, T]$ into n even intervals

$\gamma = \# \text{counts in } [0, T]$ let

$\gamma_n \sim \text{Binom}(n, \lambda/n)$ then

the # counts is expected
to be independent of n .

$$\lim_{n \rightarrow \infty} P(\gamma_n = y) = \lim_n \binom{n}{y} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y}$$

$$= \lim_n \frac{n!}{(n-y)!} \frac{1}{y!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^y \frac{\lambda^y}{n^y}$$

$$= \frac{e^{-\lambda} \lambda^y}{y!} \lim_n \frac{n!}{(n-y)! n^y}$$

$$= \frac{e^{-\lambda} \lambda^y}{y!} \lim_n \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-y+1}{n}$$

$$= \frac{e^{-\lambda} \lambda^y}{y!}$$

Then: If $\gamma_i \stackrel{\text{iid}}{\sim} \text{Pois}(\lambda_i)$ for

Thrm: If $Y_i \stackrel{\text{ind}}{\sim} \text{Pois}(\lambda_i)$ for

$$1 \leq i \leq n \text{ then } \sum_{i=1}^n Y_i \sim \text{Pois}\left(\sum_{i=1}^n \lambda_i\right)$$

Pf:

$$\begin{aligned} M_{Y_i}(t) &= \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y e^{-\lambda}}{y!} \\ &= e^{-\lambda} \sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!} = e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t - 1)} \end{aligned}$$

Now for the sum of ind R.V.

$$M_{\sum Z_i}(t) = \prod_{i=1}^n M_{Z_i}(t)$$

So in our case

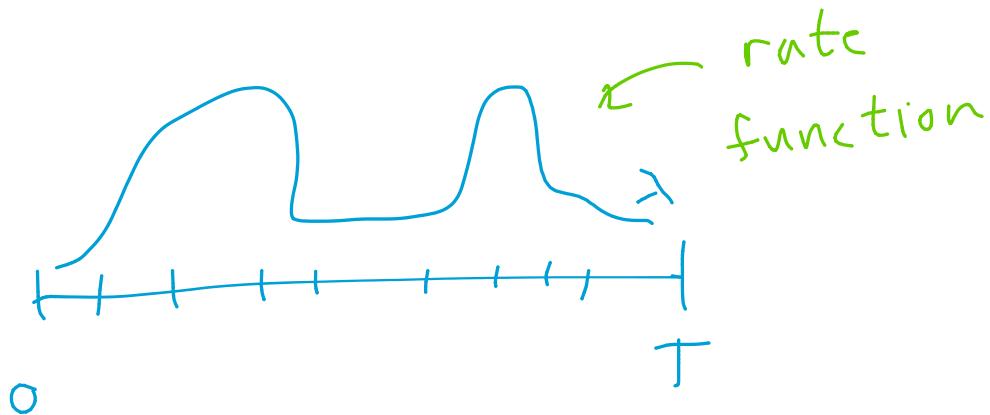
$$M_{\sum Y_i}(t) = \exp \left\{ \sum_{i=1}^n \lambda_i (e^t - 1) \right\}$$

∴ $\sum Y_i \sim \text{Pois}\left(\sum_{i=1}^n \lambda_i\right)$

$$\text{So } \sum_{i=1}^n Y_i \sim \text{Pois}\left(\sum_{i=1}^n \lambda_i\right)$$



So what happens when
 λ_i changes as some function
of an observable covariate?



$$\text{Let } \lambda(t) = \lim_{\Delta \rightarrow 0} \frac{\mathbb{E}[Y_{(t,t+\Delta)}]}{\Delta}$$

Let $Y = \# \text{ counts in } [0, T]$.

$$\text{then } Y \sim \text{Pois}\left(\sum_i \lambda(t_i) \Delta\right)$$

for a fixed partition

In the limit

$$Y \sim \text{Pois} \left(\int_0^T \lambda(t) dt \right)$$

This type of process is called an inhomogeneous

Poisson Process.

Poisson GLM

For $Y_i \sim \text{Pois}(\lambda_i)$

$$P(Y_i = y) = \exp \left\{ y \log \lambda_i - \lambda_i - \log(y!) \right\}$$

$$\theta = \log \lambda_i, \quad \lambda_i = e^\theta \Rightarrow b(\theta) = e^\theta$$

$$\mu = b'(\theta) = e^\theta = \lambda_i$$

$$\text{Var}(Y_i) = \alpha(\theta) b''(\theta) = \lambda_i$$

$$D(\hat{y}; \hat{\mu}_i) = 2 \sum_{i=1}^n y_i (\bar{\theta}_i - \hat{\theta}_i) - [b(\bar{\theta}_i) - b(\hat{\theta}_i)]$$

$$= 2 \sum_{i=1}^n \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\}$$

$$= 2 \sum_{i=1} \left\{ y_i \log \frac{\hat{\mu}_i}{\hat{\mu}_{i.}} - (y_i - \hat{\mu}_{i.}) \right\}$$

\approx Bern. \approx Normal
 Deviance Deviance

D n aym $\vec{\chi}_{n-p}^2$

The canonical link

$$g(\mu_i) = \log(\mu_i) = \theta$$

$$\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

or

$$\lambda_i = \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}$$

Lecture 2/26

Monday, February 26, 2018 1:21 PM

Overview

Last time: - Poisson Dist.
- Poisson Processes

This Time: - Poisson GLM

Reading : - M&N Chp. 5

Poisson GLM

$$y_i \sim \text{ind Pois}(\lambda_i)$$

$$\begin{aligned}\text{Canonical link } \eta_i &= \log(\mu_i) \\ &= \log(\lambda_i) \\ &= x_i^T \beta\end{aligned}$$

This implies

$$\lambda_i = e^{x_i^T \beta} \text{ for } \lambda_i \geq 0$$

Simple Model $\lambda_i = e^{\beta_0 + \beta_1 x_i}$

$$e^{\beta_0} = \mathbb{E}[Y_i | X_i = 0]$$

$$\beta_0 = \log [\mathbb{E}(Y_i | X_i = 0)]$$

"log of baseline
rate"

e^{β_1} modulation of the baseline
rate when x_i is increased
by 1.

Ex: Y_i = hypothermic cases
at ER per day

x_i = avg daily temp

Note: date occurs cont.
binned



Let's assume we just take temp to be constant over the day

GLM: Fit gives

$$\hat{\beta}_0 = \ln(3) \quad \text{expected \# cases at } 0^\circ \text{ is 3}$$

$$\hat{\beta}_1 = \ln(0.9) \quad \mathbb{E}[Y_i | x_i = 1] = (3)(0.9) < 3$$

cases decreases

as x increased

$$\begin{aligned} \mathbb{E}[Y_i | X_i = -20] \\ = (3)(.9)^{-20} = 24.7 \end{aligned}$$

cases increases
as x decreases

Aside on Binning: Much Poisson

data originates from continuous
Poisson process in time.

Sometimes time scale is set
by the experiment.

Other times the data
is available in continuous
time

Q: Factors influencing choice

of bin size?

1. Time scale of covariates
2. Time course of prediction
3. Trade off between n and $E(Y_i)$
4. Accuracy of assumptions
(i.e. independence)

Weighting by Exposure

$Y_i = \# \text{ events in } n \text{ known}$
interval t_i .

Building a model with rate

$$\lambda_i = \exp\{x_i^T \beta\}$$

$$Y_i \sim \text{Pois}(t_i; \lambda_i)$$

$$\Rightarrow \mu_i = t_i e^{x_i^T \beta} = e^{x_i^T \beta + \log t_i}$$

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \ln(\mu_i) - \ln(t_i)$$

Consider this as an extension
of the link function with
an offset $\ln(t_i)$.

- The fitting procedure is identical
- The estimates of the intercept change to

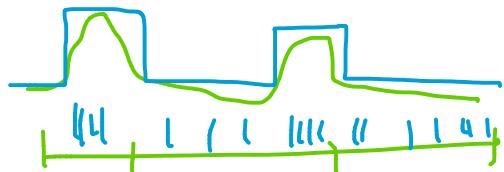
$$\mathbb{E} \left[\frac{\text{# events}}{\text{unit time}} \right]$$

- Remaining parameters identical

Ex: X_i = T arrivals at BU East

$$x_i = \mathbb{1}_{\{\text{Rush hour}\}}$$

$$t_i = \begin{cases} 4 \text{ hrs} & \text{rush hour} \\ 1.5 \text{ hrs} & \text{non rush} \end{cases}$$



GLM: $\lambda_i = e^{\beta_0 + \beta_1 x_i}$

$$\hat{\beta}_0 = \log(6)$$

off rush hour

we expect 6

trains per

hour or

84 trains

$$\hat{\beta}_1 = \log(2)$$

we expect

$$e^{\log(6)} e^{\log(2)} = 12$$

trains per hour

on rush hour

full day we expect

48 trains during

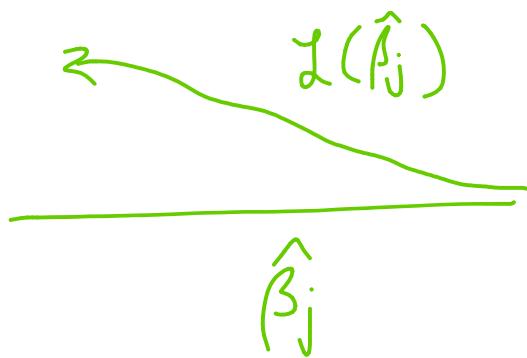
rush hour.

What can go wrong?

1. $(X^T W X)$ not invertible

2. If $y_i = 0$ for all $x_{ij} \neq 0$

then $\hat{\beta}_j = -\infty$ ($\lambda = e^{4r + \sigma_1 x}$)



Structural zeros: $\lambda_i = 0$

$x_i \neq 0 \Rightarrow y_i = 0$

Solution: $\hat{\beta}_j = -\infty$ and remove

all datapoints with $x_i = 0$

Sampling zeros: $\lambda_i > 0$ for some

$x_i \neq 0$ but $y_i = 0$

Solution: Collect more data

Often the ML fits are good
but inferences based on I_B
are incorrect.

Lecture 2/28

Wednesday, February 28, 2018 1:24 PM

Overview

Last time: Poisson GLM

This time: Over/under dispersion

Reading: M & N 6.4-6.7

Dispersion

$$Y_i \sim \text{ind Pois}(\lambda_i)$$

$$\text{Var}(Y_i) = \mathbb{E}[Y_i] = \lambda_i$$

Often $\widehat{\text{Var}}(Y_i)$ is significantly greater than λ_i - Overdispersion

Sometimes $\widehat{\text{Var}}(Y_i) < \lambda_i$ - Underdispersion.

Often evidence through

$$\chi^2 \text{ or } \{(r_p)_i\}$$

Overdispersion - sign of

poor fit.

Options: 1) improve model

2) model dispersion directly

Sources of overdispersion:

- Unmodeled variability in λ_i
- Dependence between events

1. Assume λ_i is random

$$\mathbb{E}[y_i] = \mathbb{E}[\mathbb{E}[y_i | \lambda_i]]$$

$$= \mathbb{E}[\lambda_i]$$

$$\text{Var}(y_i) = \mathbb{E}[\text{Var}(x_i | \lambda_i)] + \text{Var}(\mathbb{E}[x_i | \lambda_i])$$

$$= \mathbb{E}[\lambda_i] + \text{Var}(\lambda_i)$$

$\geq \mathbb{E}[Y_i]$.

So we have an increased variance given by $\text{Var}(\lambda_i)$

Common model $\lambda \sim \Gamma(\alpha, \beta)$

with density

$$f(\lambda) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$$

$Y_i \sim \text{Neg Binomial}(r=\alpha, p=\frac{1}{\beta+1})$

with density

$$P(Y=y) = \binom{y+r-1}{y} (1-p)^r p^y$$

for $y=0, 1, 2, \dots$

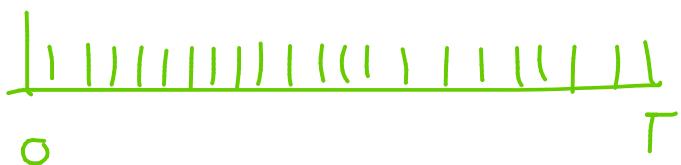
This events aren't necessarily independent.

Note: Neg. Bin. is in the
... - - - model

Note: Neg. bin. is an exp. family so we can model y_i with a neg. bin. G.L.M.

Poisson process describes independent events occurring cont. in time and space.

What if the events are dependent.



Partition interval s.t.

$y_i \sim \text{Pois}(\lambda/n)$ marginally.
but $\text{Corr}(y_i, y_j) = \rho_{ij}$.

$$\text{Var}\left(\sum^n y_i\right) = \sum^n \text{Var}(y_i) + ? \sum_{i,j} \text{Corr}(y_i, y_j)$$

$$\text{Var}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \text{Var}(Y_i) + 2 \sum_{i \neq j} \text{Cov}(Y_i, Y_j)$$

$$= \lambda + 2 \sum_{i \neq j} \underbrace{\beta_{ij} \sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}}_{\text{can be neg. or pos. so}}$$

can be neg.

or pos. so

this term

can lead to
over/under dispersion.

Model the joint distribution

$$P(Y_1, \dots, Y_n) = \prod_{i=1}^n P(Y_i | Y_1, Y_2, \dots, Y_{i-1})$$

We can construct these probabilities directly.

Let $H_i = (Y_{i-1}, Y_{i-2}, \dots, Y_1)$.

$$Y_i | H_i; X_i \sim \text{Pois}(\lambda(t_i | X_i))$$

$$\text{where } \lambda(t_i | X_i) = E[Y_i | H_i]$$

This function is called the
conditional intensity.

So we can still model
this using a Poisson GLM
but need bins small
enough to capture behavior
of $\lambda(t_i | H_i)$.

Include H_i as covariates
along with the X_i .

Most common:

1) Autoregressive Analogue.

$$\log(\lambda_i) = X_i^T \beta + \sum_{k=1}^K \gamma_k Y_{i-k}$$

Still GLM so we
can use

- Likelihood Estimation
- Consistent ML Estimates
- Asy. Normal
- F.I. includes Covariance information.

Only slight difference is
that the design matrix
is also stochastic.

2) Estimate the dispersion

Assume $\text{Var}(Y_i) = \sigma^2 \lambda_i$

and estimate σ^2 (No longer

Poisson).

$$\text{Let } \hat{\sigma}^2 = \frac{\chi^2}{n-p}$$

Then $(n-p) \frac{\hat{\sigma}^2}{\sigma^2} \xrightarrow{\text{asy}} \chi^2$

Use χ^2 CI. of hypothesis tests to estimate $\hat{\sigma}^2$.

Actual estimates of β are unaffected by $\hat{\sigma}^2$.

Only difference is Fisher information $\frac{1}{\sigma^2}(X^T W X)$

$$\Rightarrow \text{Cov}(\hat{\beta}) = \hat{\sigma}^2 (X^T W X)^{-1}$$

For inferences on β_j

$$t = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \quad \text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 (X^T W X)^{-1}_{jj}}$$

For analysis of deviance

$$F = \frac{D(y; \hat{\mu}) - D(y; \hat{\mu}_2)}{(p-q)} / \hat{\sigma}^2 \sim F_{p-q, n-p}$$

$$\frac{1}{(P-z)} \quad \cancel{\alpha^i}$$

Overview

Last time: Pois GLM examples

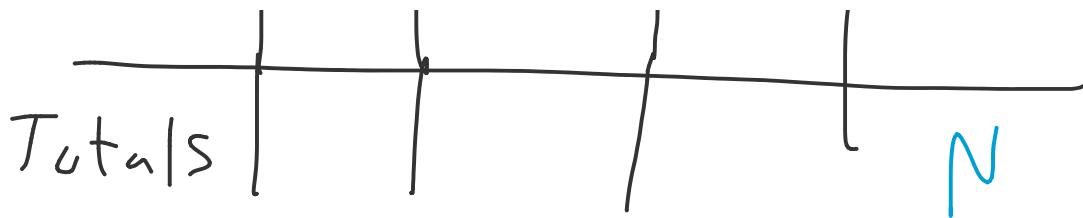
This time: Contingency tables

Note: Take homemidterm

- 4 hrs.

Contingency Tables

Factor A	a_1	\dots	a_m	Totals
Factor B	b_1	\dots	\dots	
\vdots	\vdots	\ddots	\ddots	
	b_m			



Modeling Strategy:

- depends on which is considered fixed or controlled.

Ex:

Program\Grade	A	B	C	D	F
UG					
MA					
PhD					

Strategies:

• Build a multinomial GLM
for $\text{Grade} = g(\text{Program})$

Q: What if all factors
are randomized?

Ex: Gender, eye color, hair color

M	Green	Black
F	Hazel	Brown
	Blue	Red
	Brown	Blonde

Can we detect a relation
between these variables?

Strategies

- Build 3 marginal models
taking each one fixed.

taking exam ...

- Poisson GLM
 - Consider a spatial point process across cells
 - Treat each cell as a dependent variable that depends on its factor levels.

Intuition: Continuous process of discrete events that fill up the table.

Let $y_{ij} = \# \text{ datapoints in}$

cell (i,j) . Then we model

$$Y_{ij} \sim \text{Pois}(\lambda p_{ij})$$

ground
rate

↑ [effect
of cell

where $\sum_i \sum_j p_{ij} = 1$

Let $S = \sum_i \sum_j Y_{ij} \sim \text{Pois}(\lambda)$

λ is unknown but we know

$$S = N$$

Conditional dist. of

$$Y_{ij} | S = N$$

$$P(Y_{ij} = y | S=N)$$

$$= \frac{P(Y_{ij} = y, S=N)}{P(S=N)}$$

$$= \frac{P(Y_{ij} = y, S - Y_{ij} = N-y)}{P(S=N)}$$

$$= \frac{P(Y_{ij} = y) P(S - Y_{ij} = N-y)}{P(S=N)}$$

$$= \frac{\frac{e^{-\lambda p_{ij}}}{(\lambda p_{ij})^y} \cdot \frac{e^{-\lambda(1-p_{ij})}}{(\lambda(1-p_{ij}))^{N-y}}}{y! (N-y)!}$$

$$\frac{e^{-\lambda} \lambda^N}{N!}$$

$$= \frac{N!}{y!(N-y)!} (P_{ij})^y (1-P_{ij})^{N-y}$$

$$= \binom{N}{y} (P_{ij})^y (1-P_{ij})^{N-y}$$

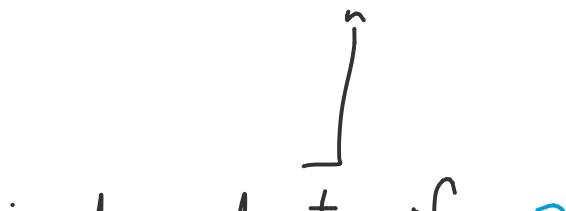
$\sim \text{Bin}(N, P_{ij})$ \leftarrow independent
of λ

So given N Binomial & .

Poisson are the same .

Similarly : joint dist of

$$\{Y_{ij}\}_{S=N} \sim \text{Multi}(N, \{P_{ij}\})$$



↓
independent of γ

So implicitly building Bernoulli
Multinomial models.

Use Poisson GLM to model.

$$\lambda_{ij} = \lambda p_{ij} \Rightarrow \log(\lambda_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}$$
$$= \log(\lambda) + \log(p_{ij})$$

• We can then interpret

the model estimates
as relative probabilities.

Ex: Gender \times hair color \times eye color
 \sim 1 - an statistics students.

$N = 592$ statistics students.

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 1_{\text{Brown hair}}$$

$$+ \beta_2 1_{\text{Black hair}} + \dots + \beta_7 1_{\text{male}}$$

Interpretation:

e^{β_0} = expected # student in
Baseline category.

e^{β_i} = modulation from baseline
category for level i .

Lecture 3/14

Wednesday, March 14, 2018 1:27 PM

Overview

Last Time: Contingency Table

This Time: Gamma GLM

Reading: M&N Chp 8.

Nonnegative R.V.

Usually for $X \geq 0$

we see $\text{Var}(X) \nearrow$ as

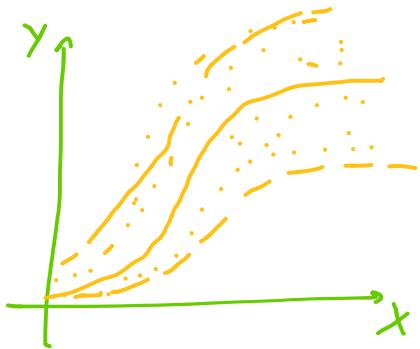
$E(X) \nearrow$ (e.g. Gamma, Poisson).

Gamma GLM Theory

The Gamma models

cont., positive data with
constant coefficient of
variation given by

$$CV = \frac{\sqrt{\text{Var}(Y_i)}}{\mathbb{E}(Y_i)}$$



Gamma pdf (standard)

$$Y \sim \Gamma(\alpha, \beta)$$

$$f_Y(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}$$

for

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

$$y > 0, \alpha > 0, \beta > 0$$

↑
Shape ↑
 scale

Rmk: For certain values of (α, β) we can get χ^2 and exponential.

Properties: $E(Y) = \alpha\beta$ $V(Y) = \alpha\beta^2$

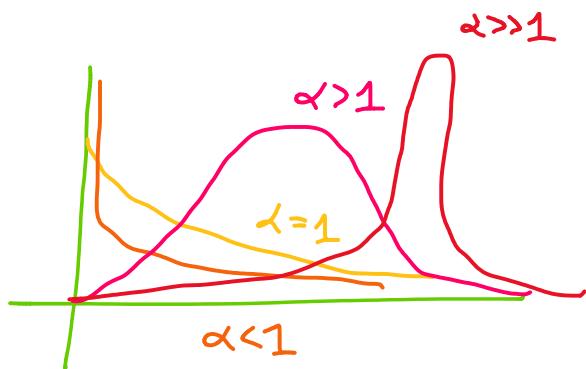
$$CV = \sqrt{\frac{1}{\alpha}}$$

So scaling between variance +
II, has \propto

mean controlled by α

Lastly

$$m_x(t) = (1 - \beta t)^{-\alpha}$$



Reparameterize to use (μ, r) . Let

$$\mu = \alpha \beta > 0 \quad r = \alpha > 0$$

$$\alpha = r \quad \beta = \frac{\mu}{r}$$

Then our parameterization is

$$m_x(t) = \frac{r^r}{u^{r-1}} e^{-\frac{y^r}{\mu}}$$

$$f_y(y) = \frac{r^{-r}}{n^r \Gamma(r)} y^{r-1} e^{-\gamma/\mu}$$

Claim: $\underline{\Gamma}$ is in the exp. family.

$$f_y(y) = \exp\left(\frac{\gamma(-y_n) - \log(n)}{y_r} + c(y, n) \right)$$

$$\theta = -\frac{1}{n}$$

$$b(\theta) = \log\left(-\frac{1}{\theta}\right) = -\log(-\theta)$$

$$a(\phi) = \frac{1}{r}$$

$$\mathbb{E}(Y) = b'(\theta) = \frac{-1}{\theta} = \mu$$

$$\text{Var}(Y) = b''(\theta) a(\phi) = \frac{1}{\theta^2 r} = \frac{n^2}{r}$$

Then the variance in terms

of the mean

$$V(\mu) = \mu^2$$

Find the canonical link

$$g \text{ s.t. } g'(\mu) = \Theta = -\frac{1}{\mu}$$

(Write Θ in terms of μ)

When we go to set

$$-\frac{1}{\mu} = \eta = X\beta \quad \text{we can}$$

just flip signs on β

So our link could

$$\text{be } \eta = \pm \frac{1}{\mu} = X\beta$$

Possible Concerns

1. $g^{-1}(x\beta) : \mathbb{R} \rightarrow \mathbb{R}$

· does not impose $\mu > 0$.

2. Interpretability

3. Extrapolation near impossible.

Ex: $Y_i = \text{value of used car}$

$X_i = \text{time owned.}$

Using Gamma GLM

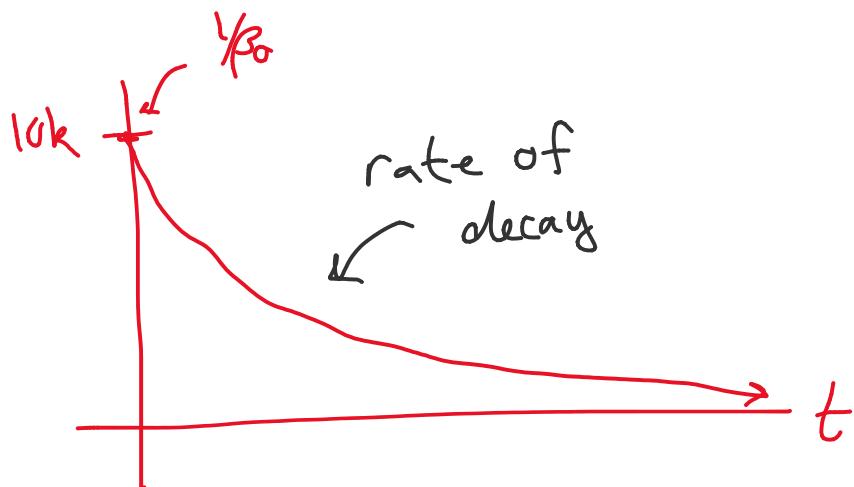
$$\frac{1}{M_i} = \beta_0 + \beta_1 X_i$$

$$\Rightarrow M_i = (\beta_0 + \beta_1 X_i)^{-1}$$

Suppose we say

$$\alpha = \frac{1}{\beta} \quad \beta = \frac{1}{\text{per year}}$$

$$\beta_0 = \frac{1}{10k} \quad \beta_1 = \frac{1}{50h \cdot \text{year}}$$



$$\text{As } \beta_1 = \frac{1}{5} \beta_0 \quad \text{so}$$

after 5 years the car
loses $\frac{1}{2}$ of its initial
Value.

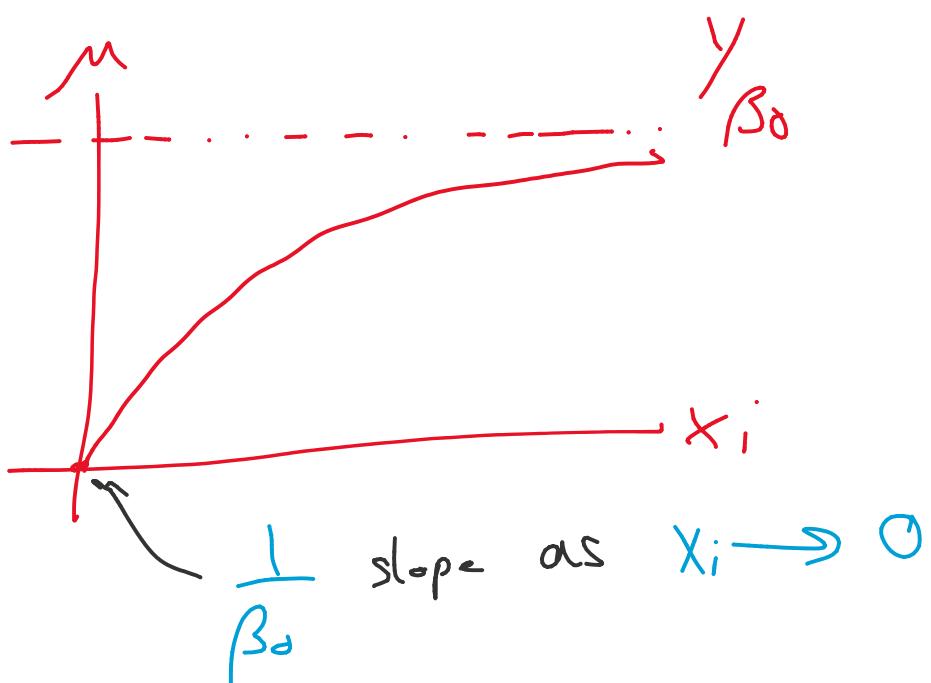
After 10 years the
car decreases it's value
by $\frac{2}{3}$.

Ex: y_i = Cost of repairing
used car. x_i = time owned.

The we want to set
up the model as

$$\frac{1}{m_i} = \beta_0 + \beta_1 \frac{1}{x_i}$$

$$m_i = \frac{x_i}{\beta_0 x_i + \beta_1}$$



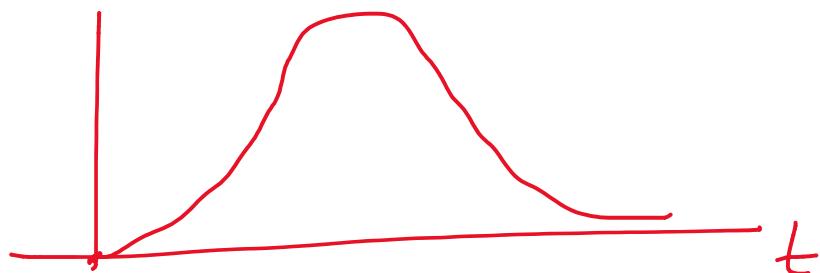
Ex: y_i = money actually spent

Ex: Y_i = money actually spent
on repairs

X_i = time owned

$$\frac{1}{M_i} = \beta_0 + \beta_1 \frac{1}{X_i} + \beta_2 X_i$$

$$M_i = \frac{X_i}{\beta_0 + \beta_1 X_i + \beta_2 X_i^2}$$



Overview

Last Time: Gamma GLM

This Time: other link functions
 Goodness of fit
 Estimating Disp.

Gamma GLM

$$Y_i \sim \text{ind} \Gamma(r, \frac{\mu_i}{r})$$

$$\mathbb{E}(Y_i) = \mu_i \quad V(Y_i) = \mu_i^2$$

Canonical link

$$g(\mu_i) = \frac{1}{\mu_i} = X_i^T \beta$$

(doesn't ensure $\hat{\mu} > 0$)

We will use other

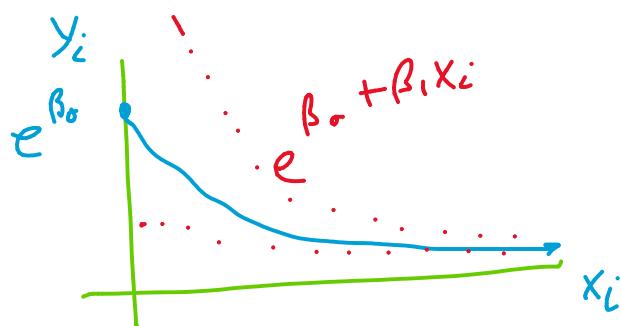
We will use other link functions s.t. $\hat{\mu} > 0$

1. $g(\mu_i) = \log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$

useful for when x_i

have mult. effects on μ_i

Ex: y_i = value of car
 x_i = time owned



If $B_0 = \log \$10k$

$B_1 = \log 0.9$

New car is worth 10k

For each year a car decreases in value by $1 - e^{-B_1} \approx 0.9$

Decreases in value
a factor of $e^{\beta_1} = 0.9$

A car loses 10% of
its value each year.

- Decay is faster than
canonical e^{-x} vs. λx

2. Identity link $\eta_i = \mu_i$

($\hat{\mu}_i$ can be negative)

Used often for modeling
sample variances and
sums of squares (χ^2 data)

Ex: y_i = sample variance for

R.S. grades

x_i = problem set #

$j = 1, 2, \dots, n_i$

$j=1, 2, \dots, r_i$

$g_{ij} = \text{grade } j^{\text{th}} \text{ student review}$
on the i^{th} problem set.

$$g_{ij} \sim N(m_i, \sigma_i^2)$$

$$Y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (g_{ij} - m_i)^2$$

$$\frac{n_i Y_i}{\sigma^2} \sim \chi^2_{n_i} = \Gamma\left(\frac{n_i}{2}, z\right)$$

$$\Rightarrow Y_i \sim \Gamma\left(\frac{n_i}{2}, \frac{z \sigma^2}{n_i}\right)$$

$\downarrow \quad \downarrow$

$$\frac{\sigma_i^2}{r}$$

r is known in this
case.

Building the GLM.

$$\eta_i = \sigma_i^2 = \beta_0 + \beta_1 X_i$$

IF $\beta_0 = 95$ $\in \mathbb{R}^1 - 1 \times 1$

If

$$\beta_0 = 95 \quad E(Y_i) = 100$$

$$\beta_1 = 5$$

Goodness of Fit

Deviance:

$$D(y_i; \hat{\mu}) = 2 \sum_{i=1}^n y_i (\tilde{\theta}_i - \hat{\theta}_i) \\ - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))$$

for

$$\tilde{\theta}_i = -\frac{1}{\mu_i} \quad b(\theta_i) = \log(\mu_i)$$

$$D(y_i; \hat{\mu}) = 2 \sum y_i \left(-\frac{1}{y_i} + \frac{1}{\hat{\mu}_i} \right) \\ - (\log(y_i) - \log(\hat{\mu}_i))$$

$$= 2 \sum_{i=1}^n \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) - \log \left(\frac{y_i}{\hat{\mu}_i} \right)$$

Note: for real gamma data

$$P(Y_i = 0) = 0$$

$$P(Y_i=0) = v$$

If $y_i=0$ and $\hat{\mu}_i \neq 0$

$$Per(y; \mu) = +\infty$$

Multiple deviance corrections

exist for datasets with

0 values

Pearson Stat. χ^2

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{v(\mu_i)} = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2}$$

Estimating Dispersion

In some cases v may be known. (χ^2 model, $v = \frac{n_i}{2}$)

When v is unknown we estimate via ML and MOM.

1. ML

$$\log L(r) = \sum_{i=1}^n \frac{y_i(-\frac{1}{m_i}) - \log(m_i)}{y_r}$$

$$+ r \log r + (r-1) \log(y_r) \\ - \log(\Gamma(r))$$

$$\frac{d \log M L}{dr} = \sum_{i=1}^n y_i(-\frac{1}{m_i}) - \log(m_i)$$

$$+ \log(r) + 1 + \log(y_r) - \frac{\Gamma'(r)}{\Gamma(r)} = 0$$

digamma $\psi(r)$

\Rightarrow

$$n(\log(r) - \psi(r)) = \sum_{i=1}^n \left(\frac{y_i - m_i}{m_i} \right) - \log \frac{y_i}{m_i}$$

$$\Rightarrow \text{Der}(y_i, \hat{\mu}) = 2n(\log(r) - \psi(r))$$

Solve for r to find MLE.

Via approximation

$$\hat{r}_{ML} = \frac{n(6n+2D)}{n(1+D)}$$

$$r_{ML} = \frac{\dots}{D(b_n + D)}$$

2. MM: Let $z_i = \frac{y_i}{m_i}$

$$\mathbb{E}(z_i) = \frac{\mathbb{E}(y_i)}{m_i} = 1$$

$$\text{Var}(z_i) = \frac{\text{Var}(y_i)}{m_i^2} = \frac{1}{r}$$

So mom

$$\frac{1}{r} = \frac{1}{(n-p)} \sum_{i=1}^n \left(\frac{y_i}{m_i} - 1 \right)^2$$

$$= \frac{\sum (y_i - m_i)^2 / m_i^2}{n-p} = \frac{\chi^2}{n-p}$$

$$\text{So } \tilde{r}_{mom} = \frac{n-p}{\chi^2}$$

Overview

Last Time: Gamma GLM links

This Time: Gamma examples,
Quasi-likelihood

Example Notes

Normal model: $\sigma^2 = \text{Var}(Y_i)$

Gamma model: $\frac{\sigma^2}{\mu} = \frac{\text{Var}(Y_i)}{\mu_i^2}$

So we have no scaling

by the mean for normal
model.

- R's deviance give a Method of Moments estimators - not a MLE.

Quasi-Likelihood Methods

- What happens when our data has an unknown distribution?
- Must it have a real dist at all?
- GLM Fitting & inference didn't make use of the full likelihood?

Instead we specified

1. How the mean depends on the data

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

2. How the variance depends on the mean.

$$\text{Var}(y_i) = \sigma^2 V_i(\mu_i)$$

3. Independence of observations

So instead of giving a distribution, just specify these three things.

If we have a known dist

$f \in \text{NEF}(\emptyset)$ then

$$\log(\mathcal{L}) = \sum \frac{y_i \theta_i - b(\theta_i)}{\alpha(\emptyset)} + c(y_i, \theta_i)$$

$$\Rightarrow \frac{d\mu_i}{d\theta_i} = b''(\theta_i) = V(m_i)$$

$$\Rightarrow U(\beta) = \sum \frac{(y_i - m_i)}{\alpha(\emptyset)V(m_i)}$$

Central Idea: Replace $U(\beta)$

with a quasi-score

$$U_i(m_i; y_o) = \frac{y_i - m_i}{\sigma^2 V(m_i)}$$

and use this as a basis
for fitting + inference.

Overview

Last Time: Gamma Examples

Intro to Quasi

This Time: Quasi GLM Est,

Quasi-Likelihood

To build a Quasi-likelihood

We need

1. Link function

2. Variance

For a single R.V. y_i

For a single R.V. Y_i

define the Quasi-Score

$$U_i = U(\mu_i; Y_i) = \frac{Y_i - \mu_i}{\sigma^2 V_i}$$

This shares a few properties with score functions.

1. $\mathbb{E}(U_i) = 0$ mom Estimator

2. $V(U_i) = \frac{\text{Var}(Y_i)}{(\sigma^2 V_i)^2} = \frac{1}{\sigma^2 V_i}$

3. $-\mathbb{E}\left(\frac{\partial U_i}{\partial \mu_i}\right)$

$$= -\mathbb{E}\left(\frac{-1}{\sigma^2 V_i} - \frac{Y_i - \mu_i}{\sigma^2 V_i} V_i'\right)$$

$$= \frac{1}{\sigma^2 V_i} = \frac{1}{\sigma^2}.$$

$$= \frac{1}{\sigma^2 v_i} = \frac{1}{\text{Var}(Y_i)}$$

Substitute Qs for

$$\frac{\partial \log L}{\partial \mu}$$

Define the quasi-likelihood

Contribution of Y_i as follows.

$$Q_i = Q(\mu_i; y_i)$$

$$= \int_{y_i}^{\mu_i} \frac{Y_i - t}{\sigma^2 v_i(t)} dt$$

(should be called quasi-
log-likelihood)

If Q_i exists it behaves

If Q_i exists it behaves like a log-likelihood

if y_i as

$$\frac{\partial Q_i}{\partial \mu_i} = u_i$$

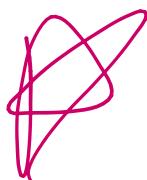
For independent data

$\{y_i\}$ the full Quasi-

Likelihood

$$Q(\vec{\mu}, \vec{y}) = \sum_{i=1}^n Q_i$$

$$= \sum_{i=1}^n \int_{y_i}^{m_i} \frac{y_i - t}{\sigma^2 v_i(t)} dt$$



Ex: Assume $V(\mu_i) = 1$

Then

$$Q = \frac{1}{\sigma^2} \sum_{i=1}^n \int_{y_i}^{\mu_i} (y - t) dt$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^n \int_0^{y_i - \mu_i} x dx$$

$$= \frac{-1}{\sigma^2} \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{2}$$

$$= \sum_{i=1}^n \frac{-(y_i - \mu_i)^2}{2 \sigma^2}$$

Ex: Assume $V(\mu_i) = \mu_i$

$$Q = \frac{1}{\sigma^2} \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - t}{t} dt$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i \log(\mu_i) - \mu_i) \Big|_{y_i}$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^n y_i \log(\mu_i) - \mu_i - y_i \log(y_i)$$

(Poisson log-likelihood)

To define a quasi-likelihood

GLM model

Let

$$\mu_i = g(\mu_i) = X_i^T \beta$$

where $g(\mu_i)$ is any invertible link function.

(No guarantees on convexity)

(No guarantees on convergence
unless we back solve)

In order to estimate β

$$\frac{\partial Q}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial Q_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \mu_i} \frac{\partial \eta_i}{\partial \beta_j}$$

$$= \sum_{i=1}^n \left(\frac{y_i - \mu_i}{\sigma^2 v_i} \right) \left(\frac{1}{g'(\mu_i)} \right) x_{ij} = 0$$

↓
Set

In matrix notation

$$\frac{dQ}{d\vec{\beta}} = X^T W \frac{(z - X\beta)}{\sigma^2}$$

$$W = \text{diag} \left((v_i g'(\mu_i)^2)^{-1} \right)$$

$$Z = \gamma - \text{diag}(g'(\mu))(\gamma - \mu)$$

In book

$$= D^T V^{-1} (\bar{\gamma} - \bar{\mu}) / \sigma^2$$

$$V = \text{diag}(V_i)$$

$$D = \begin{bmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \dots & \frac{\partial \mu_1}{\partial \beta_p} \\ \vdots & & \vdots \\ \frac{\partial \mu_n}{\partial \beta_1} & & \frac{\partial \mu_n}{\partial \beta_p} \end{bmatrix}$$

$$D_{ij} = \frac{1}{g''(\mu_i)} X_{ij}$$

Solve for $\hat{\beta}$ using IRLS.

(Quasi) Fisher Information

$$\hat{I}_{\beta} = - \mathbb{E} \left[\frac{\partial^2 Q}{\partial \beta^2} \right]$$

$$\hat{\epsilon}_{\beta}^* = - \mathbb{E} \left(\frac{\partial^2 Q}{\partial \beta \partial \beta^T} \right)$$

$$= \frac{X^T W X}{\sigma^2} = \frac{D^T V^{-1} D}{\sigma^2}$$

$$\hat{\beta} \stackrel{\text{asy}}{\sim} N(\beta, (\hat{\epsilon}_{\beta}^*)^{-1})$$

G.O.F.: Define Quasi-Deviance

$$D(\hat{\mu}, \hat{y}) = -2\sigma^2 Q(\hat{\mu}; \hat{y})$$

$$= -2 \sum_{i=1}^n \int_{y_i}^{\hat{y}_i} \frac{y_i - t}{v_i(t)} dt$$

Pearson Stat:

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{v_i} \stackrel{\text{asy}}{\sim} \sigma^2 \chi^2_{n-p}$$

$$\lambda = \sum_{i=1}^n \frac{y_i - \mu}{\sqrt{\mu}} \sim \chi_{n-p}^2$$

$$D(\hat{\mu}, \vec{y}) \xrightarrow{\text{asy}} \sigma^2 \chi_{n-p}^2$$

$$\underline{\text{Ex: }} V_i = 1, Q = -\frac{\sum (y_i - \mu_i)^2}{2\sigma^2}$$

$$\Rightarrow D(\hat{\mu}, \vec{y}) = \sum_{i=1}^n (y_i - \mu_i)^2$$

$$\underline{\text{Ex: }} V_i = \mu_i \Rightarrow$$

$$Q = \frac{1}{\sigma^2} \sum y_i \log\left(\frac{\mu_i}{y_i}\right) - (\mu_i - y_i)$$

$$\Rightarrow D(\hat{\mu}, \vec{y}) = 2 \sum y_i \log\left(\frac{\mu_i}{y_i}\right) - (\mu_i - y_i)$$

(deviance of Poisson)

Ex: $V_i = m_i^3$ = exercise

$$Q = \frac{1}{\sigma^2} \sum \frac{-y_i}{2m_i^2} + \frac{1}{m_i} - \frac{1}{2y_i}$$

$$D(m, \bar{y}) = \sum_{i=1}^n \frac{(y_i - m)^2}{y_i m_i^2}$$

Overview

Last Time: Quasi Likelihood / Deviance

This Time: Estimating σ^2

Quasi-Likelihood Methods

Quasi likelihood / deviance

$$Q(\mu, \bar{y}) = \sum_{i=1}^n \int_{y_i}^{m_i} \frac{y - t}{\sigma^2 v_i(t)} dt$$

$$D(\mu, \bar{y}) = -2\sigma^2 Q'(\mu, \bar{y})$$

Don't need to worry
about saturated model

because $D(y_{ij}\bar{y}) = 0$

To build GLM define

a link function

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Estimate $\hat{\boldsymbol{\beta}}$ using

IRLS.

in 1 ... of Fisher info.

Analogue of Fisher info.

$$l_p = \frac{\mathbf{X}^T W \mathbf{X}}{\sigma^2}, \quad W = \text{Diag} \left[V_i(\mu_i) g'(\mu_i)^{-1} \right]$$

Estimating Dispersion

1. Adjusted MOM

2. Extended Mat Quasilikelihood

$$1. \quad E \left(\frac{(Y_i - \mu_i)^2}{V_i(\mu_i)} \right)$$

$$= \frac{\text{Var}(Y_i)}{V_i(\mu_i)} = \frac{\sigma^2 V_i(\mu_i)}{V_i(\mu_0)} = \sigma^2$$

$$\bar{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n \frac{(y_i - \hat{m}_i)^2}{v_i(\mu_i)} = \frac{\chi^2}{n-p}$$

2. We need a new

notion of likelihood

because Quasi-Likelihood

maximize $c^T x$ s.t. $x^2 = \infty$

Extended Quasi-likelihood

$$Q^+(\mu, \sigma^2; \bar{y}) = Q(\mu, \bar{y}) + c_1(\sigma^2) + r_2(y)$$

To behave like a likelihood
for σ^2

$$\mathbb{E}\left(\frac{\partial Q^+}{\partial \sigma^2}\right) = \mathbb{E}\left(\frac{\partial Q}{\partial \sigma^2}\right) + c'(\sigma^2) = 0$$

Well recall that

$$Q(\mu, \bar{y}) = \frac{-D(\mu, \bar{y})}{2\sigma^2}$$

so

$$\mathbb{E}\left(\frac{\partial Q}{\partial \sigma^2}\right) + c'(\sigma^2) = 0$$

$$H(\frac{1}{\sigma^2}) + \dots$$

$$\Rightarrow c_1'(\sigma^2) = -\frac{E(D)}{2(\sigma^2)^2}$$

Last time we said

$$D \stackrel{\text{asy}}{\sim} \sigma^2 \chi_{n-p}^2$$

$$E(D) = (n-p)\sigma^2$$

So

$$c_1'(\sigma^2) = -\frac{E(D)}{2(\sigma^2)^2} \underset{\approx}{=} -\frac{(n-p)}{2\sigma^2}$$

\approx

\Rightarrow

$$C_1(\theta^2) = -\frac{(n-p)}{2} \log(\theta^2)$$

Estimate $\hat{\sigma}_{QML}^2$:

$$\frac{\partial Q^T}{\partial \theta^2} = \frac{D(\tilde{\mu}; \vec{y})}{2(\hat{\theta}^2)^2} - \frac{(n-p)}{2\hat{\theta}^2} = 0$$

$$\Rightarrow D(\tilde{\mu}; \vec{y}) = (n-p)\hat{\theta}^2$$

$$\Rightarrow \hat{\sigma}_{ML}^2 = \frac{D(\tilde{\mu}; \vec{y})}{n-p}$$

Properties:

$$\mathbb{E}(\hat{\sigma}^2) \cong \frac{(n-p)\sigma^2}{(n-p)} = \sigma^2$$

$$\begin{aligned} \text{Var}(\hat{\sigma}^2) &= \left(\frac{\sigma^2}{n-p} \right)^2 \cdot (n-p) = \\ &= \frac{2\sigma^2}{n-p} \end{aligned}$$

$$I_{\hat{\sigma}^2} = -\mathbb{E}\left(\frac{\partial^2 Q^+}{\partial(\hat{\sigma}^2)^2}\right)$$

$$= -\mathbb{E}\left(\frac{\partial}{\partial \hat{\sigma}^2} \frac{D}{2(\hat{\sigma}^2)^2} - \frac{n-p}{2\hat{\sigma}^2}\right)$$

$$= E \left(\frac{P}{(\sigma^2)^3} - \frac{n-P}{2(\sigma^2)^2} \right)$$

$$\leq \frac{(n-p)\sigma^2}{(\sigma^2)^3} - \frac{(n-p)}{2(\sigma^2)^2}$$

$$= \frac{(n-p)}{2(\sigma^2)^2} = \frac{1}{\text{Var}(\hat{\sigma}^2)}$$

In our previous discussion
on over dispersion of Poisson

$$Y_i | \lambda_i \sim \text{Pois}(\lambda_i)$$

$$\lambda_i \sim \Gamma(\alpha, \beta)$$

$$Y_i \sim \text{Neg Bin}(\alpha, \frac{1}{1+\beta})$$

Thinking about this
in a quasi-sense

$$\text{Var}(Y_i) = \text{Var}(\mathbb{E}(Y_i)) + \mathbb{E}(\text{Var}(Y_i))$$

$$= \text{Var}(\lambda_i) + \mathbb{E}(\lambda_i)$$

$$= \alpha\beta + \alpha\beta^2$$

$$= \mu_i + \frac{\mu_i^2}{\sigma} \quad \xrightarrow{\text{quad.}} \quad \begin{array}{l} \text{func. of} \\ \text{mean} \end{array}$$
$$= V(\mu_i)$$

Overview

Last time: Extended Quasi
Estimating Dispersion

This time: Joint modeling of mean
and dispersion

Note: No class Friday

Joint Modeling

Previously we wrote

$$\text{Var}(Y_i) = \phi V_i(\mu_i)$$

Covariates only influence
variance through $\hat{\mu}_i$.

More generally we could allow the variance to also depend on the covariates with its own link & dist.

$$h(\phi_i) = \sum_{j=1}^p \delta_j u_{ij}$$

Γ

invertible may be same
as X covariates

Q: How do we estimate (γ, β) together.

One Idea: Use extend Quasi-likelihood

$$Q^+(\hat{\mu}, \bar{y}) = \sum_{i=1}^n \int_{\mu_i}^{y_i} \frac{y_i - t}{\phi_i v_i(t)} dt$$

\nearrow

$$-\frac{1}{2} \log(\phi_i) + c_2(\vec{y})$$

y_i
m_i

add this)

$$D_i = -2 \int_{m_i}^{y_i} \frac{y_i - t}{v_i(t)} dt$$

Quasi-Deriance component

from the i^{th} datapoint.

$$Q^+(\mu, y) = - \sum_{i=1}^n \frac{D_i}{2\alpha_i} - \frac{1}{2} \log \phi_i + c_2(y)$$

To estimate β

$$\frac{\partial Q^+}{\partial \beta} = \sum_{i=1}^n \frac{y_i - m_i}{\phi_i v_i(\mu_i)} \frac{d\mu_i}{\partial \beta_i} = 0$$

$r_1, \dots, r_n \hat{\beta}$ will now depend

Solving for $\hat{\beta}$ will now depend on ϕ_i . We can still solve via IRLS where we know include weights W on y_{ϕ_i}

If we can estimate γ
we can use it to estimate

$$\hat{\phi}_i = h^{-1}(\sum \gamma_i u_{ij})$$

\Rightarrow we can find \hat{w} for
IRLS and compute estimated.

To estimate γ :

$$\frac{\partial Q^+}{\partial \gamma_i} = \sum_{i=1}^n \left(\frac{d_i}{2\phi_i^2} - \frac{1}{2\phi_i} \right) \frac{d\phi_i}{d\gamma_i}$$

$$= \sum_{i=1}^n / d_i - \phi_i \backslash \frac{d\phi_i}{d\gamma_i} = n$$

$$= \sum_{i=1}^n \left(\frac{d_i - \phi_i}{2\phi_i^2} \right) \frac{\partial \phi_i}{\partial \delta_i} = 0$$

Looks like GLM estimating equation with data d_i , mean ϕ_i , and $V(\phi_i) = \phi_i^2$.

Solve by IRLS with link function $h(\cdot)$.

(Gamma GLM/ Quasi GLM)

Need β to estimate ϕ^\wedge

Need ϕ^\wedge to estimate β^\wedge

- Cyclic descent type algorithms.

Fit simultaneous GLM for μ and dispersion.

mean and dispersion
using residuals as data
for estimating δ .

More generally specify
a joint model as follows:

$$\mathbb{E}(x_i) = \mu_i \quad \mu_i = g(\mu_i) = \sum_{j=1}^{p_m} x_{ij} \beta_j$$

$$\text{Var}(x_i) = \phi_i V_i(\mu_i) \quad \mathbb{E}(d_i) = \phi_i$$

$$g_i = h(\phi_i) = \sum_{j=1}^{p_v} v_{ij} \delta_j$$

$$\text{Var}(d_i) = \tau V_p(\phi_i)$$

where d_i is some statistic
that models dispersion

Common Choices:

- Squared Deviance Resid.
- Squared Pearson Resid.
- $V_0(\rho_i) = \rho_i^2$

Rmk: Model estimates are
dependent.

Overview

Last time: Joint models of μ and ϵ

This time: Empirical Variance Est.

Variance Est.

y_i mind from NEF
or Quasilikelihood but
we are uncertain about
the variance assumption

$$\text{Var}(y_i) \stackrel{?}{=} V_i(\mu_i)$$

If n is sufficiently large
estimate $\text{Cov}(\hat{\beta})$ directly
from data.

Using $V_i(\mu_i)$ we
constructed the score
function

$$U_i(\beta_j) = \frac{d \log L_i}{d \beta_j}$$

$$= \frac{d \log L_i}{d \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$$

depends
on Var *depends*
on link

$$= \frac{\partial \mu_i}{\partial \beta_j} \frac{y_i - \mu_i}{\phi V_i(\mu_i)}$$

We trust that solving
this for $\hat{\beta}$ gives good
estimates

$$U(\hat{\beta}) = \frac{\partial \log L}{\partial \beta} = \frac{1}{n} D^T V^{-1} / (\gamma_M)$$

$$U(\hat{\beta}) = \frac{\partial \log L}{\partial \beta} = \frac{1}{n} \mathbf{D}' \mathbf{v}' (\mathbf{y}_M)$$

$$\mathbf{D} = \left(\frac{\partial \mu_i}{\partial \beta_j} \right)_{i,j=1}^n$$

$$\mathbf{V} = \text{Diag}(v_i(\mu_i))$$

Create: Explore $U(\hat{\beta})$ near $\hat{\beta}_{MLE}$ to get some idea of variance

Taylor series expansion of $U(\hat{\beta})$ about β

$$U(\hat{\beta}) = U(\beta) + (\hat{\beta} - \beta) \frac{\partial U(\beta)}{\partial \beta} + O(\beta)$$

Setting to zero

$$\begin{aligned} (\hat{\beta} - \beta) &\simeq -U(\beta) \left(\frac{\partial U(\beta)}{\partial \beta} \right)^{-1} \\ &\simeq U(\beta) \mathbb{E} \left(-\frac{\partial U(\beta)}{\partial \beta} \right)^{-1} \end{aligned}$$

$$= u(\beta) \mathbb{E} \left(-\frac{\partial u(\beta)}{\partial \beta} \right)$$

$$= u(\beta) \underbrace{(I(\beta))^{-1}}_{\text{Notion of Covariance}}$$

$$\mathbb{E} \left(-\frac{\partial u(\beta)}{\partial \beta} \right)$$

$$= \mathbb{E} \left(-\frac{\partial u(\beta)}{\partial \mu} \frac{\partial \mu}{\partial \beta} \right)$$

$$= \frac{1}{\phi} \mathbb{E} \left(D^T V^{-1} D \right) + \frac{1}{\phi} \mathbb{E} \left(D^T \frac{\partial u^{-1}}{\partial \mu} (y - \mu) D \right)$$

$$= \frac{1}{\phi} \mathbb{E}_y \left(D^T V^{-1} D \right)$$

$$= \frac{D^T V^{-1} D}{\phi}$$

$$\mathbb{E}(u(\beta)) = \frac{1}{\phi} \mathbb{E}(D^T V^{-1} (y - \mu))$$

$$\mathbb{E}(u(\beta)) = \frac{1}{\phi} \#(D \setminus \{y = 0\}) \\ = 0$$

So

$$(\beta - \tilde{\beta}) \stackrel{\text{asy}}{\sim} N(0, \sigma^2)$$

$$\hat{\beta} \stackrel{\text{asy}}{\sim} \text{unbiased}$$

$$\text{Var}(u(\beta)) = \frac{1}{\phi} \mathbb{E} \left[D^T V^{-1} (y - \mu) (y - \mu)^T V^{-1} D \right]$$

$$= \frac{D^T V^{-1} \text{Cov}(Y_i) V^{-1} D}{\phi^2}$$

If our model was
specified as $\text{Var}(Y_i) = \phi V$

$$\text{Var}(u(\beta)) = \frac{D^T V^{-1} D}{\phi} \quad \checkmark$$

$$= \frac{A}{\phi^2} \quad \text{Can't reduce further.}$$

Now define $J = D^T V^{-1} D$

All of this together

$$(\hat{\beta} - \beta) \stackrel{\text{asy}}{\sim} N(0, \frac{A}{\phi^2}) \left(\frac{D^T V^{-1} D}{\phi} \right)^{-1}$$

$$\stackrel{\text{asy}}{\sim} N(0, \hat{\Sigma})$$

$$\hat{\Sigma} = J^{-1} A J^{-1} \quad \hat{\beta}$$

Robust Estimator

Empirical Estimator

Sandwich Estimator

If we want to solve

for $\hat{\beta} \cdot J^{-1}$ is covariance

$C \approx$ under $\text{Var}(Y) = \phi V_i / n_i$

of $\hat{\beta}$ under $\text{Var}(Y) = \phi V_i(\mu_i)$
 We correct with A .

Estimate

$$\text{Cov}(Y) = \text{Diag} \left[(y_i - \hat{\mu}_i)^2 \right]$$

$$\Rightarrow V^{-1} \text{Cov}(Y) \approx \text{Diag} \left(r_{p_i}^2 \right)$$

\nearrow
Pearson

From here we can write

$$\hat{A} - D V^{-1} \widehat{\text{Cov}}(Y) V^{-1} D = X^T W^* X$$

$$W^* = \text{Diag} \left(\frac{r_{p_i}^2}{g'(\mu_i)^2 V_i} \right)$$

Weight is now a function

of link, variance and Pearson
residuals.

Base tests & inference
for this corrected model.

Lecture 4/2

Monday, April 2, 2018 11:12 AM

Overview

Last time: Empirical variance

This time: Survival Analysis

Intro to Survival

Let y_1, \dots, y_n be a

sequence of nonneg. R.V.

representing waiting times.

Problem: Model relation

for (y_i, \bar{x})

Challenge: At any fixed

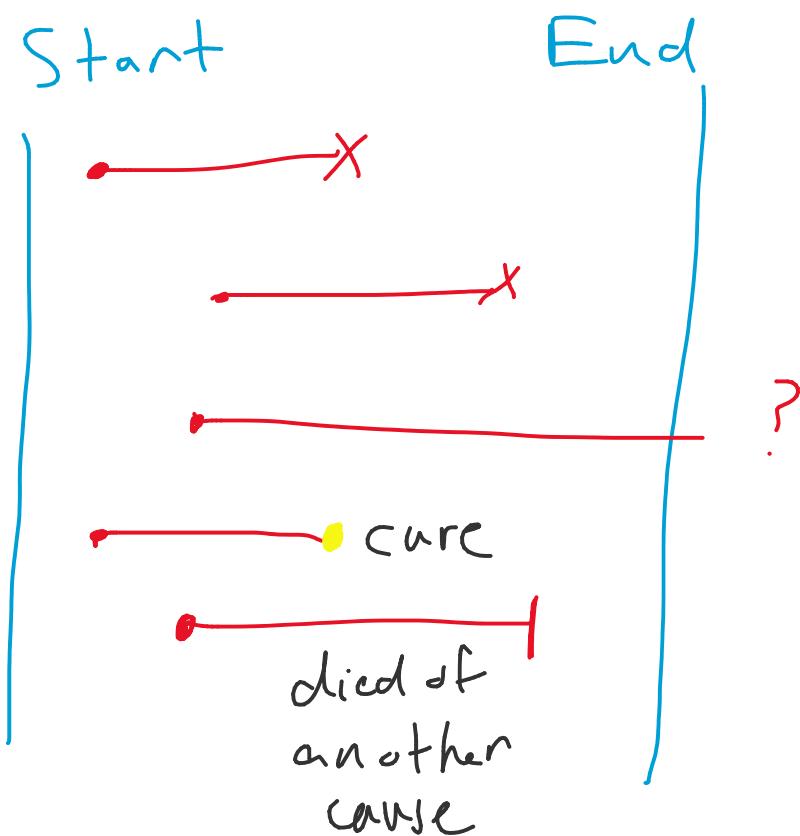
experiment end time,

outcomes (death v. no death)

outcomes (death v. no death)
of some unit will not
be known

- Censored data
- Missing, not random

Ex: Cancer Study



Throwing out data

Throwing out data

not really an option

- Want to analyze effect on survival & death
- Leads to bias estimates and reduces power

Instead of observing Y_i

(censored data point) we

only observe some feature

e.g. $Y_i \geq y$.

Modeling Survival Data:

Assume $Y_i \sim \text{iid } f(t)$

Cont.

, . , | | | | | ... $F(t)$

with distribution $F(t)$

We'll make some regularity assumptions on $f(t)$

Def: Survival function

$$S(t) = P(Y \geq t) = 1 - F(t)$$

So

$$f(t) = -\frac{dS(t)}{dt}$$

From here the mean survival time can be derived as

$$\begin{aligned} E(Y_i) &= \int_0^\infty tf(t)dt \\ &= \int_0^\infty [1 - S(t)] dt \end{aligned}$$

] By parts

$$= \int_0^\infty S(t) dt \quad \boxed{\text{Pai}}$$

Def: The hazard function

is given by

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq Y_i \leq t + \Delta t | Y_i > t)}{\Delta t}$$

Serves as a notion of instantaneous risk of death at time t

$$\underline{h(t)} = \lim_{\Delta t \rightarrow 0} \frac{P(t < Y_i \leq t + \Delta t)}{\Delta t P(Y_i > t)}$$

$$= \frac{1}{P(Y_i > t)} \lim_{\Delta t \rightarrow 0} \frac{P(t < Y_i \leq t + \Delta t)}{\Delta t}$$

$$= \frac{1}{S(t)} \quad \partial \frac{F(t)}{\partial t}$$

$S(t)$

$$= \frac{f(t)}{S(t)} = -\frac{d \log(S(t))}{dt}$$

Then the cumulative hazard

$$\underline{H(t) = \int_0^t h(u)du}$$

$$= -\log(S(t)) + \log(S(0))$$

$$= -\log(S(t))$$

Another way to think
of our survival function

is

$$-H(t) = -\int_0^t h(u)du$$

$$S(t) = e^{-H(t)}$$

This gives a 1-1 relation

This gives us '...' between the hazard & survival function

\Rightarrow we can build our model from either.

$$f(t) = - \frac{ds(t)}{dt}$$

$$= h(t) e^{- \int_0^t h(u) du}$$

$$= h(t) S(t)$$

\int [prob of making it to time t

hazard of dying

when you get there.

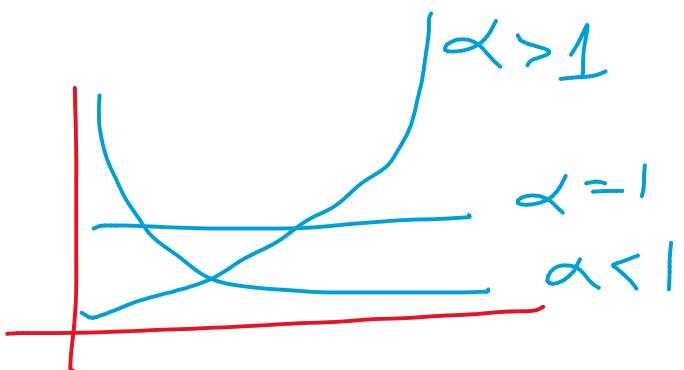
Ex: Constant hazard

$$h(t) = \lambda \text{ then } S(t) = e^{-\lambda t}$$

$$f(t) = \lambda e^{-\lambda t} \sim \text{Exp}(\lambda)$$

$\Rightarrow Y_i$ are the waiting times of a homogeneous Poisson process.

Ex: $h(t) = \lambda \alpha t^{\alpha-1} (\alpha, \lambda > 0)$



$$S(t) = e^{-\lambda t^\alpha}$$

$$f(t) = \lambda \alpha t^{\alpha-1} e^{-\lambda t^\alpha}$$

$$f(t) = \lambda \alpha t^{\alpha-1}$$

$$\Rightarrow Y_i \sim \text{Weibull}(\lambda, \alpha)$$

Censored Likelihood

Assume we have n uncensored datapoints and m censored data points

$$(Y_1, \dots, Y_n) \quad (Z_{n+1}, \dots, Z_{n+m})$$

where $\{Z_i = z\} = \{Y_i > z\}$

$$L = \prod_{i=1}^n h(Y_i) S(Y_i) \prod_{j=n+1}^{n+m} S(Z_j)$$

Uncensored *Censored*

— In survival analysis

-hazard only applied
to observed deaths.

$$\ell = \sum_{i=1}^n \log h(y_i) + \log S(y_i)$$
$$+ \sum_{i=n+1}^m \log S(z_i)$$

define

$$w_i = \begin{cases} 1 & i \text{ uncensored} \\ 0 & i \text{ censored} \end{cases}$$

$$\ell = \sum_{i=1}^{n+m} w_i \log h(t_i) + \log l(t_i)$$

for $t_i = \begin{cases} y_i & w_i = 1 \\ z_i & w_i = 0 \end{cases}$

Lecture 4/4

Wednesday, April 4, 2018 1:26 PM

Overview,

Last time: Intro to Survival

This time: Proportional Hazard

Survival Analysis

Recall w_i indicated whether

i obs. is censored and t_i be the time of the event or censoring of i th obs.

$$\log L = \sum_{i=1}^n w_i \log(h(t_i)) + \log S(t_i)$$

Ex: $h(t) = \lambda \Rightarrow$

$$S(t) = e^{-\lambda t} \quad Y_i \sim \text{Exp}(\lambda)$$

$$\partial_{\lambda} L = \sum w_i \log \lambda - \lambda t_i$$

$$\log L = \sum w_i \log \lambda - \gamma t_i$$

$$\frac{\partial \log L}{\partial \lambda} = \frac{\sum w_i}{\lambda} - \sum t_i = 0$$

$$\Rightarrow \hat{\lambda}_{MLE} = \frac{\sum w_i}{\sum t_i}$$

Total # uncensored
 Cumulative exposure time

Rmk: $\hat{\lambda}_{MLE} < \frac{\# \text{ uncensored events}}{\text{exposure of uncensored.}}$

$$\frac{\partial^2 \log L}{\partial \lambda^2} = -\frac{\sum w_i}{\lambda^2}$$

$$\Rightarrow I_{\hat{\lambda}_{MLE}} = \frac{\sum w_i}{\hat{\lambda}_{MLE}^2} = \frac{(\sum t_i)^2}{\sum w_i}$$

Obs. fisher

So under reg. conditions.

So under reg. conditions.

$$\hat{\lambda} \stackrel{\text{asy}}{\sim} N\left(\lambda, \frac{\sum w_i}{(\sum t_i)^2}\right)$$

Ex: $h(t) = \alpha \lambda t_i^{\alpha-1}$

$$S(t) = e^{-\lambda t_i^\alpha}$$

$$Y_i \sim \text{Weibull}(\alpha, \lambda)$$

$$\log L = \sum w_i \log \alpha + w_i \log \lambda$$

$$+ w_i (\alpha-1) \log t_i - \lambda t_i^\alpha$$

$$\frac{\partial \log L}{\partial \lambda} = \frac{\sum w_i}{\lambda} - \sum t_i^\alpha = 0$$

$$\Rightarrow \hat{\lambda}_{MLE} = \frac{\sum t_i^{\hat{\alpha}_{MLE}}}{\sum w_i}$$

$$\frac{\partial \log L}{\partial \alpha} = \frac{\sum w_i}{\alpha} + \sum w_i \log t_i$$

$$-\sum \lambda \log(t_i) t_i^\alpha = 0$$

No analytic solution.

Again we don't need
a full probability model here.

Proportional Hazard Models

Fundamental Assumption : Dependence on (\vec{x}, t_i) is mult.
separable.

$$h(t; \vec{x}_i) = \lambda(t) \exp\{\vec{x}_i^T \beta\}$$

From this we can build up
the model

$$H(t; \vec{x}_i) = \int_0^t \lambda(u) e^{m_i} du$$

$$= e^{m_i} \Delta(t)$$

$$S(t; \vec{x}_i) = e^{-\Lambda(t; \vec{x}_i)}$$

$$= \exp \left\{ -\Lambda(t) e^{m_i} \right\}$$

Proportional hazard log-likelihood

$$\log L = \sum_{i=1}^n w_i \log \lambda(t) + w_i m_i - \Lambda(t) e^{m_i}$$

from here we can argue
its in the N.E.F. if

$$m_i = \Lambda(t_i) e^{m_i}$$

$$\Rightarrow m_i = \log m_i - \log \Lambda(t_i)$$

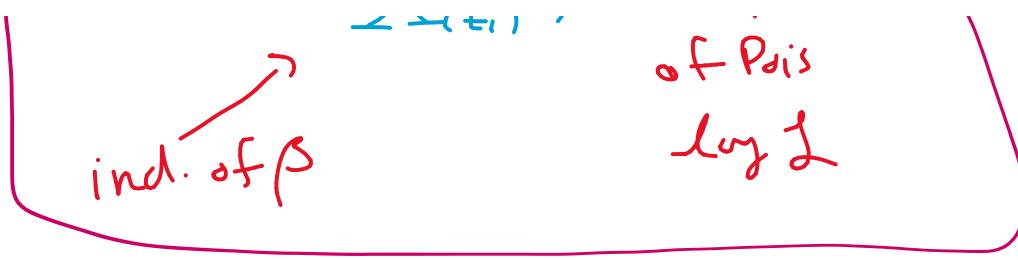
Serves as a link. Now plugging

back into log-like.

$$\log L(\beta) = \sum (w_i \log m_i - m_i)$$

$$+ \sum \left(w_i \log \frac{\lambda(t_i)}{\Lambda(t_i)} \right)$$

Component
of Pairs



key point:

For known $\lambda(t)$ we can fit

this using the poisson GLM.

with log link and offset
given by $\log \lambda(t_i)$.

Interpretation of $\hat{\beta}$

$$\text{Let's say } m_i = \beta_0 + \beta_1 x_i$$

$$h(t; x_i) \Rightarrow \lambda(t) e^{\beta_0 + \beta_1 x_i}$$

$$= e^{\beta_0 + \log \lambda(t) + \beta_1 x_i}$$

\int
Offset on
intercept

$$h(t; x=0) = e^{\beta_0 + \log \lambda(t)}$$

$$\Rightarrow \beta_0 = \log \frac{h(t; 0)}{\lambda(t)}$$

$$\Rightarrow \beta_0 + \log - \lambda(t)$$

log hazard ratio

Note: If $\beta_0 = 0$ then

$\lambda(t)$ is the baseline hazard.

β_0 = scaling on $\lambda(t)$

Consider

$$\frac{h(t; x_i = x+1)}{h(t; x_i = x)} = \frac{\lambda(t) e^{\beta_0 + \beta_1(x+1)}}{\lambda(t) e^{\beta_0 + \beta_1 x}}$$

$$\Rightarrow \beta_1 = \log \frac{h(t; x_i = x+1)}{h(t; x_i = x)}$$

how much more likely

to die for each increase
in x

Ex: $\lambda(t) = \lambda$.

$$\ln(1 + \frac{1}{\lambda}) \sim x_i^\top \beta$$

$$h(t; \vec{x}_i, \cdot) = \lambda e^{x_i^T \beta}$$

Goal: estimate (β, λ)

$$\underline{\Delta}(t) = \lambda t \quad m_i = \lambda t e^{x_i^T \beta}$$

$$\log \underline{L}(\beta, \lambda) = \sum w_i \log t_i + w_i \log \lambda \\ + w_i \log(t_i) + w_i x_i^T \beta - \lambda t e^{x_i^T \beta}$$

Estimate $\hat{\beta}$ using IRLS

(Poisson GLM)

$$\frac{\partial \log(\underline{L}(\beta, \lambda))}{\partial \lambda} = \sum_{i=1}^n w_i - \sum_{i=1}^n t_i e^{x_i^T \beta}$$

$$\Rightarrow \hat{\lambda}_{MLE} = \frac{\sum w_i}{\sum t_i e^{x_i^T \hat{\beta}}}$$

$$\text{Ex: } \lambda(t) = \lambda t^{\alpha-1}$$

(homework)

$$\sum_{i=1}^n \left\{ (\alpha_{-i}) w_i \log(t) + w_i \log(\lambda_\alpha) + w_i x_i^T \beta - \lambda t e^{x_i^T \beta} \right\}$$

Lecture 4/6

Friday, April 6, 2018 1:21 PM

Overview

Last time: Prop. Hazards

This time: Prop. Odds Model

Prop. Likelihood

Censored Survival Data

w_1, \dots, w_n = event indicators

t_1, \dots, t_n = death/time or censoring

$$\log L = \sum_{i=1}^n w_i \log h(t_i) + \log g(t_i)$$

From here we built the
Proportional Hazards Model.

$$h(t; \vec{x}_i) = \lambda(t) e^{m_i}$$

$$\log L = \sum_{i=1}^n w_i \log m_i - m_i + w_i \log \left\{ \frac{\lambda(t_i)}{\dots} \right\}$$

$$+ w_i \log \left\{ \frac{\lambda(t_i)}{\Delta(t_i)} \right\}$$

ind of β

Fit with Poisson GLM

with $\log \Delta(t_i)$ offsets

where

$$\Delta(t) = \int_0^t \lambda(u) du$$

$$m_i = \exp(\log \Delta(t_i) + m_i)$$

Proportional Odds Model

The odds of
dying/event before t
are given by

$$\frac{F(t)}{1 - F(t)} = \frac{F(t)}{S(t)} = \Lambda(t) e^{m_i}$$

where $\Lambda(\cdot)$ is nondecreasing.

So rearranging we have

$$S(t) = \frac{1}{1 + \Lambda(t) e^{m_i}}$$

To find the hazard we have

$$-\log S(t) = \log (1 + \lambda(t)e^{m_i})$$

$$h(t) = -\frac{\log S(t)}{dt}$$

$$= \frac{-\lambda'(t)e^{m_i}}{1 + \lambda(t)e^{m_i}}$$

$$= \frac{\lambda'(t)}{\lambda(t)} - \frac{\lambda(t)e^{m_i}}{1 + \lambda(t)e^{m_i}}$$

From here we can calculate the log-likelihood as follows.

$$\log L = \sum w_i \log \frac{\lambda(t_i)e^{m_i}}{1 + \lambda(t_i)e^{m_i}}$$

$$+ \log \left(\frac{1}{1 + \lambda(t_i)e^{m_i}} \right)$$

$$+ w_i \log \left(\frac{-\lambda'(t_i)}{\lambda(t_i)} \right)$$

offset

Now here we can write

$$\Theta_i = \log \frac{\lambda(t_i) e^{m_i}}{1 + \lambda(t_i) e^{m_i}}$$

$$1 - e^{\Theta_i} = \frac{1}{1 + \lambda(t_i) e^{m_i}}$$

and

$$b(\theta) = -\log(1 - e^{\theta})$$

$$a(\theta) = 1$$

$$m_i = b'(\Theta_i)$$

$$= \frac{e^{\Theta_i}}{1 - e^{\Theta_i}} = \lambda(t_i) e^{m_i}$$

$$b''(\theta_i) = \frac{e^{\theta_i}}{1 - e^{\theta_i}} + \frac{e^{2\theta_i}}{(1 - e^{\theta_i})^2}$$

$$= m_i + m_i^2$$

which is the neg. binomial.

Estimate β with

Geom/Neg-Binom/Quasi GLM

with log link and

log $\lambda(t_i)$ offset.

or II r l.h.s. models

log(λ(t))

Both of these models
rely on the fact

$\lambda(t), \Delta(t)$ are mult.

Separable from m_i .

If $\Delta(t)$ known - use
as offset.

$\Delta(t)$ unknown - construct
a parametric model for

$\lambda(t)$ and fit iteratively.

(e.g. $\lambda(t) = \lambda_0 t^{\alpha-1}$)

OR we can treat

$\lambda(t)$ as a nuisance
parameter and estimate

β using partial likelihood.

OR

Treat survival analysis
as a point process
with additional temporal

Corariates.

Partial Likelihood

For each time we observe an event, what is the likelihood it came from the i th unit.

$$\text{Let } I_j(t_i) = \begin{cases} 1 & t_j \geq t_i \\ 0 & t_j < t_i \end{cases}$$

where t_j = time of j th event.

Let the risk set be

$$R(t_i) = \{j : I_j(t_i) = 1\}$$

with this we define the partial likelihood as

$$\text{PL}_i(\beta) = P(i \text{ dies at } t_i | \text{An element of } R(t_i) \text{ dies at } t_i)$$

$$= \frac{h(t_i; x_i)}{\sum_{j \in R(t_i)} h(t_i; x_j)}$$

Risk of i

Cum. Risk

Then for the prop. hazard

model

$$PL_i(\beta) = \frac{\lambda(t_i)e^{M_i}}{\sum_{j=1}^n I_j(t_i)\lambda(t_i)e^{M_i}}$$

$$= \frac{e^{M_i}}{\sum_{j=1}^n I_j(t_i)e^{M_i}} \xrightarrow{\text{independent of nuisance}} \lambda(t)$$

So the full partial likelihood

is

$$PL(\beta) = \prod_{i=1}^n [PL_i(\beta)]^{w_i}$$

$$\log PL(\beta) = \sum_{i=1}^n w_i \log PL_i(\beta)$$

$$= \sum_{i=1}^n \left\{ w_i M_i - w_i \log \sum_{j=1}^n I_j(t_i) e^{M_i} \right\}$$

Ex: Data:

x_i	2	2	1	3
t_i	1	(6)	7	15

Censored

then for the hazard

model

$$h(t, x_i) = \lambda(t) e^{x_i \beta}$$

$$PL(\beta) = \frac{e^{2\beta}}{e^{2\beta} + e^{2\beta} t + e^{2\beta} t^2 + e^{3\beta}}$$

$$\times \frac{e^\beta}{e^\beta + e^{3\beta}} \times \frac{e^{3\beta}}{e^{3\beta}}$$

Lecture 4/9

Monday, April 9, 2018 1:22 PM

Overview

Last time:

- Prop. Odds surv.

- Partial likelihood

This time:

- Point Process

Survival

- Fitting example

Reading - Circular data

Proportional Hazards

Model the hazard directly by

$$h(t) = \lambda(t) \exp\{x^T \beta\}$$

Fit with Poisson GLM

log-link, $\log \lambda(t)$ offset

Proportional Odds

$$\frac{F(t)}{S(t)} = \lambda(t) e^{x_i^T \beta}$$

Fit with geometric
GLM, log-link,
 $\log \lambda(t)$ offset.

Generalizing

If $\lambda(t)$ unknown

- (i) Fit a parametric model to estimate $\lambda(t)$ and solve iteratively
- (ii) Partial likelihood
- (iii) General point process survival model

Point Process GLM

We'll be building models for the full hazard function

$$h(t; x_i) = \exp \left(\beta_0 + \sum_{j=1}^p \beta_j g_j(t, x_i) \right)$$

arbitrary

arbitrary
functions

- GAM with log-link

Approach: Partition time into small intervals and treat events using P.P. likelihood

Note: if bins are small then the point process likelihood is well approximated by Poisson likelihood.

Original data description:

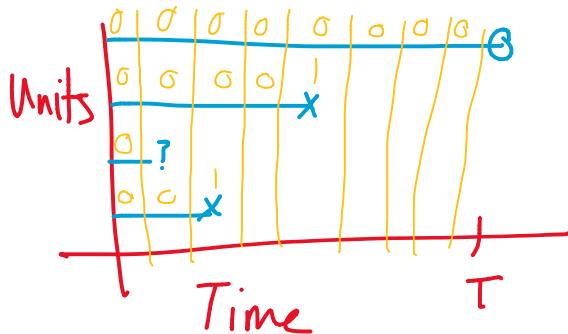
$$w_i = \begin{cases} 1 & \text{uncensored} \\ 0 & \text{censored} \end{cases}$$

t_i = times of events/
censoring.

New data will look like

$$\Delta N_{ijk} = \begin{cases} 1 & \text{event occurs for } i \text{ in } (t_k, t_{k+1}) \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \dots, n$
 $k = 1, \dots, n_i$



O - alive

X - dead

? - Last contact

0/1 - value of $\Delta N_{i,k}$

Now, we model

$$N_{i,k} = h(t_k; x_i) \Delta t$$

$$= \exp \left\{ \sum_{j=1}^n \beta_j g_j(t_k; x_i) \right\}$$

$$\log L = \sum_{i=1}^n \sum_{k=1}^{n_i} \left\{ \Delta N_{i,k} \log(h(t_k; x_i)) - h(t_k; x_i) \Delta t \right\}$$

which we fit with a
 Poisson GLM (or Binomial
 GLM for small Δt)

Notes:

... , " , . , k

Notes:

- Δt should be small enough

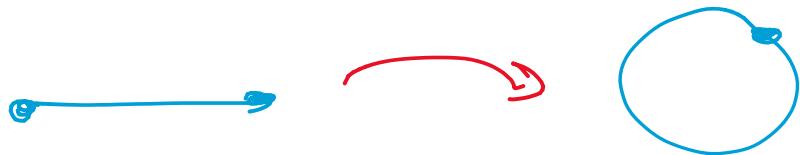
s.t. $\mathbb{E}(\Delta N_{ik}) \ll 1$

- To compare deviance/AIC
with other models we
need to use the same
data (e.g. (w_i, t_i) (ΔN_{ik}))

Lecture 4/11

Friday, May 11, 2018 8:44 AM

Circular Data



Let τ_i be directional

data on the interval

$$(-\pi, \pi]$$

Here, the behaviour at

$$\{\pi + 2\pi k : k \in \mathbb{Z}\}$$

should be consistent.

Our methods should
be independent of the
branch cuts.

Transform the data

such that

$$\left\{ \underset{1 \leq i \leq n}{:} \begin{pmatrix} c_i \\ s_i \end{pmatrix} = \begin{pmatrix} \cos(\epsilon_i) \\ \sin(\epsilon_i) \end{pmatrix} \right\}$$

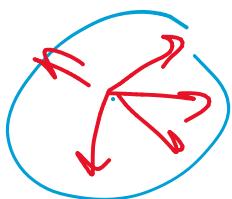
All data now lies
on the unit circle.

If we have time

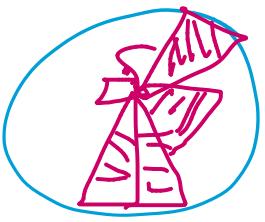
data

$$\left\{ \underset{1 \leq i \leq n}{:} \begin{pmatrix} c_i \\ s_i \end{pmatrix} = \begin{pmatrix} \cos(2\pi w t_i) \\ \sin(2\pi w t_i) \end{pmatrix} \right\}$$

Visualization



→ Raw circular

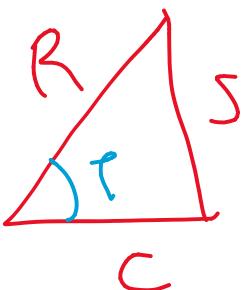


→ Rose plot
→ angular hist.

Moments

Define

$$C = \sum_i \cos(\theta_i)$$



$$S = \sum_i \sin(\theta_i)$$

$$R = \sqrt{S^2 + C^2}$$

$$\bar{R} = \frac{1}{n} R$$

$\bar{R} \downarrow$ lower precision more variable

$\bar{R} \uparrow$ more precision less variable.

Variable.

$$\bar{\ell} = \text{atan} 2(s, c)$$

mean angle for
resultant vector.

- Central tendency

- $\text{atan} 2 = \tan^{-1}\left(\frac{y}{x}\right)$

- Bigger magnitude
 \Rightarrow vectors point in
same dir.

Def: The circular variance

is given by

$$\hat{\sigma}^2 = 1 - \bar{R}$$

$$\hat{V} = 1 - \bar{R}$$

$$C \leq \hat{V} \leq I$$




least variable. most variable

Def: The circular s.f.

clv is given by

$$\hat{\sigma} = \sqrt{-2 \log R}$$

$$C \leq \hat{\sigma} \leq \infty$$

J L
last most var

Lecture 4/13

Friday, April 13, 2018 1:23 PM

Overview

Last time: Intro to Circular

This time: Linear - circular
associations

Von Mises Dist.

Circular Data

$$\{ \varphi_i \in [0, 2\pi) \text{ endpoints identified} \}$$

or

$$\{ (\zeta_i, s_i) = (\cos(\varphi_i), \sin(\varphi_i)) \}$$

$$\subseteq \{ (x, y) : x^2 + y^2 = 1 \}$$

Linear - circular association:

- Predict a variable

$y_i \in \mathbb{R}$ as a function

of a circular variable

ℓ_i

Use GLM for y_i

$$\text{with } g(\mu_i) = \beta_0 + \sum_{j=1}^p \beta_j g_j(\ell_i)$$

To handle the circular structure of ℓ_i we typically use trigonometric basis functions $\{g_j(\ell_i)\}$

Ex: y_i = # home runs in a baseball game

x_i = wind speed

ℓ_i = wind direction

$$y_i \sim \text{Pois}(\lambda_i)$$

Setting

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i \cos(\ell_i - \ell_p)$$

"Best" wind

Cond. for hitter.

Parameters : $\{\beta_0, \beta_1, \ell_p\}$

wouldn't
be linear

S_o

$$\log \lambda_i = \beta_0 + \beta_1 x_i \cos \ell_i \cos \ell_p \\ + \beta_1 x_i \sin(\ell_i) \sin(\ell_p)$$

$$\text{Let } \gamma_1 = \beta_1 \cos(\ell_p)$$

$$\gamma_2 = \beta_1 \sin(\ell_p)$$

then

$$\log \lambda_i = \beta_0 + \gamma_1 x_i \cos \ell_i$$

$$+ \gamma_2 x_i \sin \phi_i$$

Can be used for GLM
fitting.

Intercept: $e^{\beta_0} = \mathbb{E}(HR | x_i = 0)$

$$e^{\beta_1} = e^{(\gamma_1^2 + \gamma_2)^{1/2}} = \text{modulation}$$

due to wind

$$\phi_p = \alpha \tan^{-1}(r_1, r_2)$$

Circular/Linear or Circular/Circular

Associations

Goal: Predict a circular
variable ϕ_i from $\{x_i\}$
either linear or circular

Probability Models for C.D.

$$f(\varphi) : f^{(k)}(\varphi + 2\pi) = f^{(k)}(\varphi)$$

Theoretical Trig Moments:

Express data from unit vector (c_i, s_i) in \mathbb{C}

$$e^{i\varphi} = \cos(\varphi) + i\sin(\varphi)$$

Let the t.t.m

$$m_p = \mathbb{E}[(e^{i\varphi})^p]$$

$$= \int_0^{2\pi} e^{ip\varphi} f(\varphi) d\varphi$$

$$= \int_0^{2\pi} [\cos(p\varphi) f(\varphi) d\varphi + \int_0^{2\pi} \sin(p\varphi) f(\varphi) d\varphi]$$

$$= S_p e^{imp}$$

$$= f_p e^{i m_p}$$

$f = \text{mod}(m_p) = \text{resultant length}$
 $m_i = \text{expected dir.}$

Note: m_p need not be

on the unit circle

Note: $m_i \rightarrow \text{central tendency}$
 $f_i \rightarrow \text{variability}$

Let

$$v = 1 - f_i \quad (\text{circ var})$$

$$\sigma = (-2 \log f_i)^{1/2} \quad (\text{circ std. dev.})$$

Two Common Models

1. Circular Uniform:

$$f(\theta) = \frac{1}{2\pi}$$

$$m_1 = 0 + 0i$$

$$\Rightarrow \rho_1 = 0, r=1, \phi=\infty$$

$m_1 = \text{undefined.}$

- Primarily used as null hypothesis.

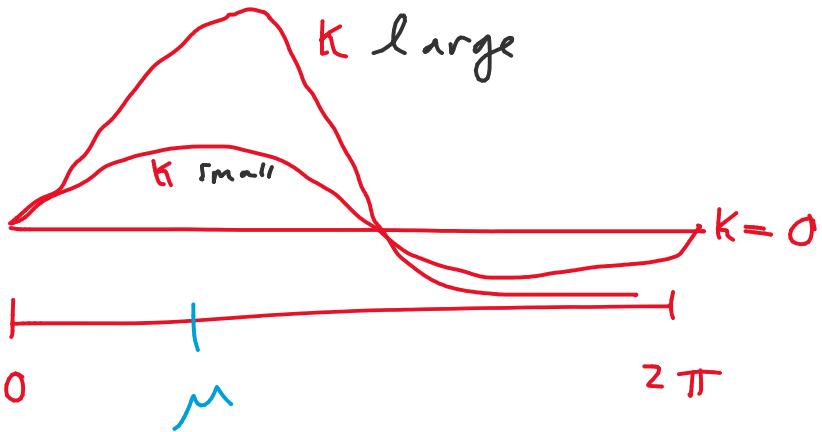
2. von Mises distribution

$$\phi_i \sim VM(\mu, k) \quad (k > 0)$$

$$f(\theta) = \frac{\exp\{k \cos(\theta - \mu)\}}{2\pi I_o(k)}$$

and

$$I_o(k) = \frac{1}{2\pi} \int_0^{2\pi} e^{k \cos(\psi)} d\psi$$



Exercise: Is VM in the
exp. family?

Lecture 4/18

Wednesday, April 18, 2018 1:20 PM

Overview

Last time: Circular data dist.

This time: von Mises GLM

von Mises

Suppose we have a dataset

$$\{ \varphi_i \in [0, 2\pi) \}$$

=

$$\{ (\zeta_i, s_i) = (\cos(\varphi_i), \sin(\varphi_i)) \in S^1 \}$$

Von Mises Likelihood:

$$L = \exp \left\{ \sum_{i=1}^n K \cos(\varphi_i - \mu_i) - \log I_0(K) + C \right\}$$

$$= \exp \left\{ \sum_{i=1}^n K (\cos \varphi_i \cos \mu_i + \sin \varphi_i \sin \mu_i) \right\}$$

$$= \exp \left\{ \sum_{i=1}^n K(\cos \varphi_i; \cos \mu_i + \sin \varphi_i; \sin \mu_i) \right. \\ \left. - \log I_0(K) + c \right\}$$

$$\boxed{= \exp \left\{ \sum_{i=1}^n \begin{pmatrix} \cos \varphi_i \\ \sin \varphi_i \end{pmatrix}^T \begin{pmatrix} K \cos \mu_i \\ K \sin \mu_i \end{pmatrix} \right. \\ \left. - \log I_0(K) + c \right\}}$$

Bivariate exponential family
with data $\begin{pmatrix} \varphi_i \\ \mu_i \end{pmatrix}$ with
natural parameter

$$\vec{\theta} = \begin{pmatrix} K \cos \mu \\ K \sin \mu \end{pmatrix}$$

$$\mu = \tan^{-1} \left(\frac{\theta_2}{\theta_1} \right)$$

$$1. \quad 1 - \gamma_1 \gamma_2$$

$$k = (\theta^T \theta)^{1/2}$$

$$b(\vec{\theta}) = \log I_o(k)$$

$$= \log I_o((\theta^T \theta)^{1/2})$$

$$a(\rho) = 1$$

Trigonometric moments

$$m_i = \frac{d b(\theta)}{dk} = \frac{I_o'(k)}{I_o(k)}$$

$$= \frac{I_o'((\theta^T \theta)^{1/2})}{I_o((\theta^T \theta)^{1/2})} \cdot \frac{2\theta}{2(\theta^T \theta)^{1/2}}$$

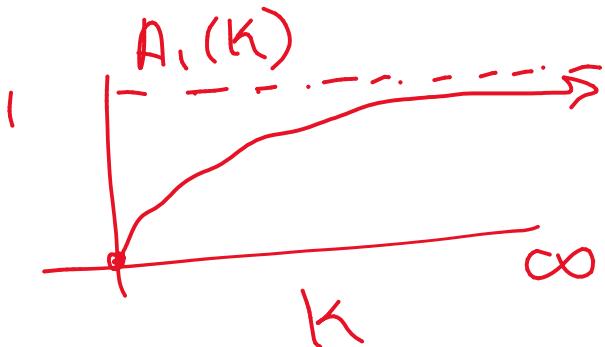
$$= A_i(k) \frac{\theta}{k}$$

$$= A_1(k) \begin{pmatrix} \cos \mu \\ \sin \mu \end{pmatrix}$$

So $m_i \in \mathbb{R}$ has length

$f_i = A_1(k)$ and direction

$$m_i = M.$$

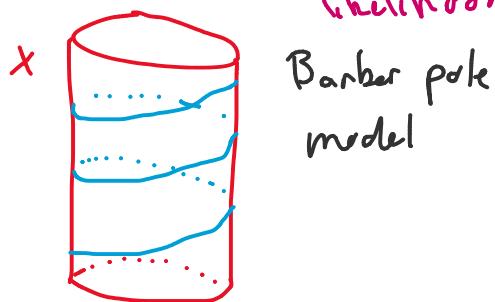


Common link functions

1. Identity $\eta_i = \mu_i \Rightarrow$

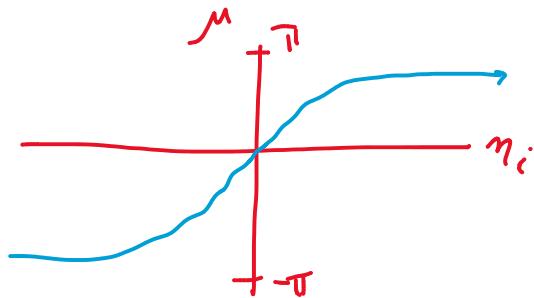
$$f(\epsilon_i) \propto e^{k \cos(-\epsilon_i - x_i^T \beta)}$$

from VM
likelihood.



$$2. \text{ Tangent link: } m_i = \tan\left(\frac{\eta_i}{2}\right)$$

$$\Rightarrow \eta_i = 2 \tan^{-1}(m_i)$$



- Maps $\mathbb{R} \mapsto [-\pi, \pi]$

Parameter Estimation

Let θ_i ~ $VM(m_0 + g(\eta_i), k)$



MCRN
link

$$J = \exp \left\{ \sum_{i=1}^n k \cos (\theta_i - m_0 - g(\eta_i)) - \log (I_r(k)) + \right\}$$

Parameters to estimation

$$(\vec{\beta}, \mu_0, K)$$

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^n K \sin(e_i - \mu_0 - g(m_i)) g'(m_i) x_{ij}$$

$$\text{let } z_i = m_i + \frac{\sin(e_i - \mu_0 - g(m_i))}{g'(m_i)}.$$

then we have

$$= \sum_{i=1}^n K (z_i - m_i) g''(m_i)^2 x_{ij} = 0$$

In matrix notation

$$X^T G^2 (\bar{z} - X \hat{\beta}) = 0$$

$$\text{for } G = \text{diag}(g'(m_i))$$

So

$$\hat{\beta} = (X^T G^2 X)^{-1} X^T G^2 \bar{z}$$

\Rightarrow

$$\hat{\beta} = (X^T G^2 X)^{-1} X^T G^2 z$$

Notice that $\hat{\beta}_i \not\in$

depends on β . So we via

IRLS.

Rmk: $\hat{\beta}$ ind. of K

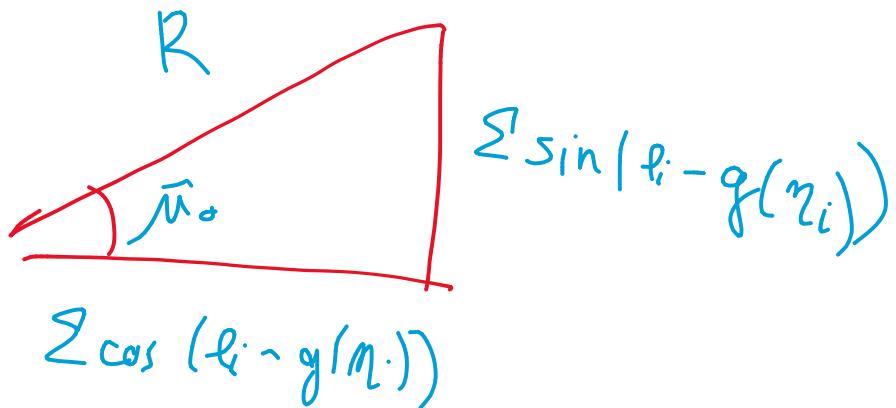
$$\frac{\partial \log L}{\partial \mu_0} = \sum_{i=1}^n k \sin(\ell_i - g(m_i))$$

$$= k \sum_{i=1}^n \sin(\ell_i - g(m_i)) \cos(\mu_0) \\ - \cos(\ell_i - g(m_i)) \sin(\mu_0) = 0$$

$$\hat{\mu}_0 = \tan^{-1} \left(\frac{\sum_{i=1}^n \sin(\ell_i - g(m_i))}{\sum \cos(\ell_i - g(m_i))} \right)$$

Rmk: Can be solved directly if

$\vec{\beta}$ is known. Use plugging in



$$R = \sqrt{\left(\sum \sin(\theta_i - g(m_i)) \right)^2 + \left(\sum \cos(\theta_i - g(m_i)) \right)^2}$$

$$\frac{\partial \log L}{\partial \kappa} = \sum_{i=1}^n \left(\cos(\theta_i - \mu_0 - g(m_i)) - \frac{I_o'(\kappa)}{I_o(\kappa)} \right) = 0$$

So

$$\cos \hat{\mu}_0 \sum \cos(\theta_i - g(m_i))$$

$$+ \sin \hat{\mu}_0 \sum \sin(\theta_i - g(m_i))$$

$$+ \sin \theta_0 < \sin(\pi - g(m_i))$$

$$= n A_i(k)$$

But from the picture...

$$\frac{(\sum \cos(\pi))^2}{R} + \frac{(\sum \sin \theta)^2}{R}$$

$$= \frac{R^2}{R} = R$$

residual
resultant
length.

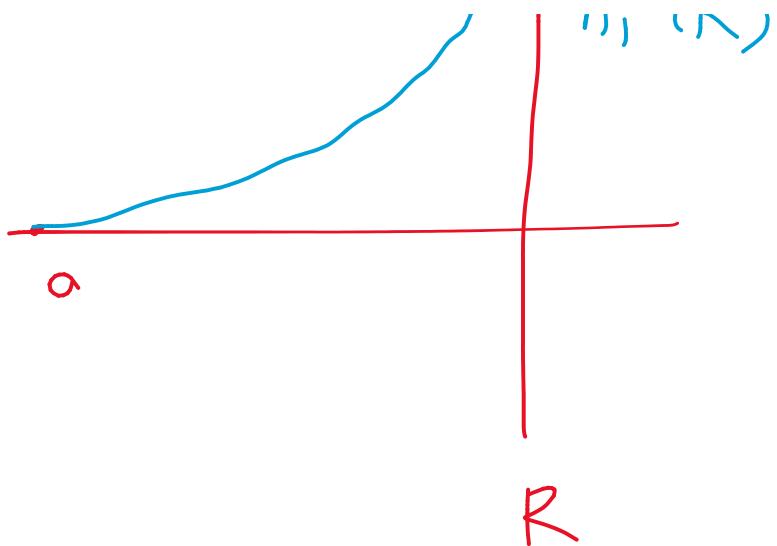
S_c

$$K^{-1} = A_i^{-1}(\bar{R}) \text{ where}$$

$$\bar{R} = \frac{R}{n} \text{ is the mean}$$

residual resultant length

$$A_i^{-1}(R)$$



Lecture 4/20

Friday, April 20, 2018 1:25 PM

Overview

Last time: Von Mises GLM

This time: Fisher Inform.

Fisher Information

Circular data:

Suppose that

$$\tau_i \sim VM(\mu_0 + g(m_i), k)$$

then

$$L(\beta, \mu_0, k) = \exp \left\{ \sum_{i=1}^n k \cos(\tau_i - \mu_0 - g(m_i)) \right. \\ \left. - \log I_0(k) + c \right\}$$

from which we can
build parameter estimates.

build parameter estimates.

Using the cumulant function
we found the first trig.
moment

$$m_i = A_i(\kappa) \begin{pmatrix} \cos \mu \\ \sin \mu \end{pmatrix}$$

Also we found parameter
estimates

$$\hat{\beta} = (X^T G^2 X)^{-1} X^T G^2 Z$$

$$G^2 = \text{diag}(g'(m_i))$$

$$Z = \eta_i + \frac{\sin(\theta_i - \mu)}{g'(\mu)}$$

which we can solve via

IRLS for

$$\hat{\mu} \Rightarrow \tan^{-1} \left(\frac{\sum \sin(\theta_i - \mu)}{\sum \cos(\theta_i - \mu)} \right)$$

We also found

$$\hat{K} = A_1^{-1}(\bar{R}) \text{ when}$$

$$\begin{aligned}\bar{R} = \frac{1}{n} \left\{ & \left(\sum_{i=1}^n \cos(\varphi_i - \mu_i) \right)^2 \\ & + \left(\sum_{i=1}^n \sin(\varphi_i - \mu_i) \right)^2 \right\}\end{aligned}$$

From which we see

$$(\hat{\beta}, \hat{\mu}_0) \perp\!\!\!\perp \hat{K}.$$

We now want to understand variability
⇒ Fisher Inform.

Preliminary Facts

$$\cdot E(\cos(\varphi_i)) = A_1(K) \cos(\mu_0 + g(\mu_i))$$

$$\mathbb{E}(\sin(\rho_i)) = A_1(K) \sin(\mu_0 + g(m_i))$$

\Rightarrow

$$\mathbb{E}[\cos(\ell_i - \mu_0 - g(m_i))]$$

$$= \mathbb{E}[\cos(\rho_i) \cos(\mu_0 + g(m_i))]$$

$$+ \sin(\ell_i) \sin(\mu_0 + g(m_i))]$$

$$= A_1(K) [\cos^2(\mu_0 + g(m_i)) + \sin^2(\mu_0 + g(m_i))]$$

$$= A_1(K)$$

and we also have

$$\mathbb{E}[\sin(\ell_i - \mu_0 - g(m_i))]$$

$$= \mathbb{E}[\sin(\rho_i) \cos(\mu_0 + g(m_i)) - \cos(\ell_i) \sin(\mu_0 + g(m_i))]$$

$$= 0$$

Fisher Information

$$\begin{aligned} & -\mathbb{E} \left[\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} \right] \\ = & \sum_{i=1}^n \left\{ K \mathbb{E} [\cos(\epsilon_i - \mu_0 - g(m_i))] \right. \\ & \quad \left. g'(m_i)^2 x_{ij} x_{ik} \right. \\ & \quad \left. - K \mathbb{E} [\sin(\epsilon_i - \mu_0 - g(m_i))] \right. \\ & \quad \left. g'(m_i)^2 x_{ij} x_{ik} \right\} \end{aligned}$$

So in matrix notation

$$\begin{aligned} & -\mathbb{E} \left(\frac{\partial^2 \log L}{\partial \beta \partial \beta^T} \right) \\ = & K A_1(K) (X^T G^2 X) \end{aligned}$$

Now with respect to μ_0^2

$$- \mathbb{E} \left[\frac{\partial^2 \log L}{\partial \mu_i^2} \right]$$

$$= \sum_k \mathbb{E} (\cos(\ell_i - \mu_k - g(m)))$$

$$= n A_1(k)$$

and with respect to K

$$- \mathbb{E} \left(\frac{\partial^2 \log L}{\partial K^2} \right) = \sum A_1'(k)$$

$$= n A_1'(k)$$

Now wrt (β_i, ζ)

$$- \mathbb{E} \left(\frac{\partial^2 \log L}{\partial \beta_j \partial K} \right)$$

$$= \sum_{l=1}^n \mathbb{E} (\sin(\ell_i - \mu_l - g(m)) g \dots)$$

$= 0$

The second derivative
w.r.t. (μ, κ) is the

same.

$$- \mathbb{E} \left(\frac{\partial \log L}{\partial \mu \partial \kappa} \right) = 0$$

And finally w.r.t (β, μ_0)

$$- \mathbb{E} \left[\frac{\partial \log L}{\partial \mu \partial \beta} \right]$$

$$= \sum K \mathbb{E} (\cos(\ell_i - \mu_0 - g(m)))$$

$$g'(m) x_{ij}$$

\Rightarrow

$$- \mathbb{E} \left[\frac{\partial \log L}{\partial \vec{\beta} \partial \vec{\mu}} \right] = K A_i(K) X^T g$$

So the full information
is $A(K) \dots A(K)$

So the full information

$$\text{matrix } \underline{\theta} = (\beta, \mu_0, K)$$

$$I(\theta) = \begin{bmatrix} P \\ A & b & C \\ b^T & C & O \\ C & O & d \end{bmatrix}$$

for $A = I_\beta = K A_1(K) (K^T a' x)$

$$b = I_{\beta, \mu_0} = K A_1(K) x^T g$$

$$C = I_{\mu_0} = n K A_1(K)$$

$$d = I_K = n A_1'(K)$$

From here we can

find the covariance

$$\text{Cov}(\theta) = I_\theta^{-1}$$

$$= \left[m A^{-1} + A^{-1} b b^T A^{-1} - A^{-1} b \quad 0 \right]$$

$$= \frac{1}{m} \begin{bmatrix} mA^{-1} + A^{-1}bb^TA^{-1} - A^{-1}b & 0 \\ -b^TA^{-1} & I & 0 \\ 0 & 0 & \frac{m}{d} \end{bmatrix}$$

for $m = c - b^T A^{-1} b$

then

$$\text{Cov}(\hat{\beta}) = \frac{1}{KA_1(K)} \left\{ 1 + \right.$$

$$\left. \frac{1}{h - g^T X (X^T A^{-1} X)^{-1} X^T g} \right\}$$

$$\text{Var}(\mu_0) = \frac{1}{h - g^T X (X^T A^{-1} X)^{-1} X^T g}$$

$$\text{Var}(k) = \frac{1}{nA_1'(K)}$$

We can then write
 the deviance as follow
 for known K

$$D(y_i, \hat{\mu}) = \frac{2}{K} \sum \begin{pmatrix} c_i \\ s_i \end{pmatrix}^T \begin{pmatrix} K_{c,i} \\ K_{s,i} \end{pmatrix}$$

$$= - \begin{pmatrix} c_i \\ s_i \end{pmatrix}^T \begin{pmatrix} K_{as\hat{\mu}_i} \\ K_{sin\hat{\mu}_i} \end{pmatrix}$$

$$= - \left(\log I_o(K) - \log I_o(\hat{K}) \right)$$

$$= 2 \sum_{i=1}^n (1 - \cos(e_i - \hat{\mu}_i))$$

Circular deviance

Lecture 4/23

Monday, April 23, 2018 1:22 PM

Overview

Last time: F.I. for VM

This time: Circular data

Circular Data

Examples :

$$\beta = (0.1, 0.65)$$

interpretation

β_0 : expected angle when

$$x=0 \quad \text{is} \quad 0.1$$

β_1 : for every 1 unit increase

in x we expect the
angle to increase 0.65

degrees.

Overview

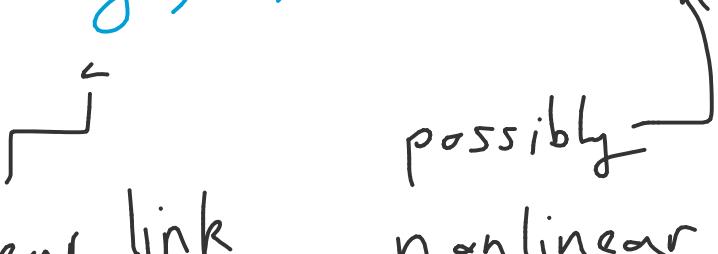
Last time: Regression Ex.

This time: Nonlinear Reg.

Nonlinear Regression

- GLMs are nonlinear by design

$$\text{• } \eta_i = g(\mu_i) = \sum \beta_j h_j(x_i)$$



What about additional

parameters in the
link, covariates, and variance
function?

e.g. $\sum_{j=1}^P \beta_j h_j(\vec{x}_i, \theta_j)$

Some approaches:

1. Recreate linear model
by Analytic linearization
2. Iterative over a low
order linear approximation
3. Fit high order linear
approximation (basis expansion)
4. Force : numerical M.L.

Analytic Linearization

Goal: relate $Y_i = \# \text{ bacteria}$

to $X_i = \text{incubation temp}$

Suppose $Y_i \sim \text{Pois}(\mu_i)$

and $\mu_i = \alpha + \beta_0 X_i + \frac{\beta_1}{2\sigma^2} X_i^2 - \frac{(X_i - \bar{X})^2}{2\sigma^2}$

Estimate $(\alpha, \bar{X}, \sigma^2)$

Notice that

$$\log(\mu_i) = \log(\alpha) + \underbrace{\frac{-m^2}{2\sigma^2}}_{\beta_0} + \underbrace{\frac{m}{\sigma^2} X_i}_{\beta_1} + \underbrace{\frac{1}{2\sigma^2} X_i^2}_{\beta_2}$$

Linear in (X_i, X_i^2)

Iteration over low order

linearization

y_i = amount of radiation

t_i = time of rad. exposure

$$y_i \sim \Gamma\left(r, \frac{m_i}{r}\right) \quad m_i = \beta (1 - e^{-\lambda t_i})$$

Assume that we have

$$m_i = \beta h(x_i, \theta) \text{ with}$$

(β, θ) unknown

If we knew θ then

we could fit β by

I.R.L.S. Start with

a guess $\theta = \theta_0$.

And use Taylor expansion

of

$$y_i \approx \beta h(x_i, \theta_0) + \gamma(\theta - \theta_0) \frac{dh}{d\theta} \Big|_{\theta=\theta_0}$$

We'll fit β and γ .

If $\theta - \theta_0$ is small then

$\hat{\beta}$ should be close to

$\tilde{\beta}_{ML}$. This corresponds to

$$\frac{\gamma}{\gamma} = (\theta - \theta_0)$$

$$\frac{\hat{\theta}}{\beta} = (\theta - \theta_0)$$

$\Rightarrow \frac{\hat{\theta}}{\hat{\beta}}$ is an accurate estimate of θ_0

If $\frac{\hat{\theta}}{\hat{\beta}}$ is large then

estimate $\theta_1 = \theta_0 + \frac{\hat{\theta}}{\hat{\beta}}$

Provides estimates of both

(β, θ) as well as

$$\text{Cov}(\hat{\beta}, \hat{\theta})$$

Then in this example

set $\lambda = \lambda_0$ then

$$m_i \approx \beta (1 - e^{\lambda_0 t_i}) + \beta (\lambda - \lambda_0) t_i e^{\lambda_0 t_i}$$

$$m_i \approx \beta \underbrace{(1 - e^{\gamma_0 x_i})}_{x_i} + \beta \underbrace{(\gamma \gamma_d)}_{\gamma} t_i \underbrace{e^{\gamma_0 x_i}}_{x_{z_i}}$$

High Order Expansion

- Expand any nonlinear components using some sort of basis expansion

$$m_i = \beta h(x_i, \theta)$$

$$= \beta \sum_{j=1}^{\infty} \alpha_j b_j(x_i)$$

↓ ↓
 parameters orthogonal
 to the model basis functions

$$\approx \beta \sum_{j=1}^d \alpha_j b_j(x_i)$$

$$= \beta \sum_{j=1}^r \alpha_j b_j(x)$$

So for the radiation uptake model

$$M_i = \beta \left(1 - \left(1 - \lambda t_i + \frac{(\lambda t_i)^2}{2!} - \frac{(\lambda t_i)^3}{3!}, \dots \right) \right)$$

$$\approx \beta \left(\lambda t_i - \frac{(\lambda t_i)^2}{2} + \frac{(\lambda t_i)^3}{3!} \right)$$

$$\approx \underbrace{\beta \lambda}_{\alpha_1} t_i - \underbrace{\frac{\beta \lambda^2}{2}}_{\alpha_2} t_i^2 + \underbrace{\frac{\beta \lambda^3}{3!}}_{\alpha_3} t_i^3$$

Rmk: more parameters

than the original model

Drawbacks: · selecting model order

- Overfitting
- Effect of outliers
- Extrapolation error.

Brute Force ML

- Numerically maximize
the likelihood

Overview

Last time: Nonlinear parameters

This time: examples

Reading: Bayesian GLM

Nonlinear Examples

- Check out Cardinal splines.

Overview

Last time: Non linear

This time: Bayesian GLM

Bayesian & Freq. Inference

Frey view of GLM:

- Model parameters represent fixed but unknown quantities.
- Estimate parameters by determining values that

make the data likely.

Bayesian view:

- Prob statements about confidence
- Uncertainty about parameters can be expressed as a subjective probability.
- Once we observe data, use Bayes' Rule to update belief.

Bayes Rule: likelihood, Prior

So if we've already run our experiment

$f(y)$ is constant so

Posterior \propto Likelihood · Prior

Specifically for GLM's

$$L_f(\beta) =$$

$$\exp\left\{\sum \frac{y_i \theta_i - b(\theta_i)}{a(\alpha)} + l(y_i, \theta)\right\} f(\beta)$$

Major issue: Choosing Priors

1. Based on belief
2. Uninformative Prior
3. Simplify calculations.

If our dataset is

sufficiently large, asym.

lead to simple solutions.

$$\hat{\beta}_{ML} \stackrel{\text{asy}}{\sim} N(\beta, I_B^{-1})$$

moreover we know

$\hat{\beta}_{ML}$ is a sufficient stat.
for β . Therefore

$$f(\hat{\beta} | \vec{y}) = f(\vec{\beta} | \hat{\beta}_{ML})$$

$$\propto f(\hat{\beta}_{ML} | \vec{\beta}) f(\vec{\beta})$$

So if prior is normal

then the posterior is

$$\text{normal}. \quad \vec{\beta} \sim N(\beta_0, \Sigma_0)$$

Then

$$f(\vec{\beta} | \vec{y}) \propto$$

$$T(P|y) \propto$$

$$\exp \left\{ -\frac{1}{2} (\hat{\beta}_{mL} - \vec{\beta})^T (I_{\vec{\beta}}^{-1})^{-1} (\hat{\beta}_{mL} - \vec{\beta}) \right\}$$

$$\times \exp \left\{ -\frac{1}{2} (\vec{\beta} - \beta_+)^T (\Sigma)^{-1} (\vec{\beta} - \beta_+) \right\}$$

$$= \exp \left\{ -\frac{1}{2} \left(\beta^T \underbrace{(\Sigma_o^{-1} + I_B)}_{\Sigma^{-1}} \beta \right. \right.$$

$$\left. \left. - \beta^T \underbrace{\Sigma^{-1} \left(\Sigma_o^{-1} \beta_o + I_B \hat{\beta}_{mL} \right)}_{\vec{\beta}} \right) \right\}$$

$$- \tilde{\beta}^T \Sigma^{-1} \beta + C \right\}$$

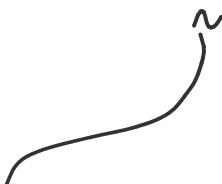
$\Sigma_o \dots$

$$\tilde{\beta} | \tilde{y} \sim N(\tilde{\beta}, \tilde{\Sigma})$$

where

$$\tilde{\Sigma} = (\Sigma_0^{-1} + I_B)^{-1}$$

$$\tilde{\beta} = \tilde{\Sigma} (\Sigma_0^{-1} \beta_0 + I_B \hat{\beta}_{ML})$$



Weighted average from
prior & data.

As $n \rightarrow \infty$ we let

then $I_B \gg \Sigma_0^{-1}$ so

our data will eventually
take over all inference.
as

$$\hat{\Sigma} \xrightarrow{n \rightarrow \infty} I_B^{-1}$$

$$\hat{\beta} \xrightarrow{n \rightarrow \infty} \beta_{ML}$$

Ex: Visual identification
experiment.

$$Y_i = \begin{cases} 1 & \text{correct} \\ 0 & \text{incorrect} \end{cases}$$

X_i = visual presentation
time

$$Y_i \sim \text{Bern}(p_i)$$

$$\text{logit}(p_i) = \beta x_i$$

and the ML estimation

$$\text{leads to } \hat{\beta}_{\text{ML}} = \mathbb{S}^{-1} I_B^{-1}$$

Now consider the Bayes App.

$$\beta \sim N(\beta_0, \Sigma_0)$$

Prior 1: $\beta \sim N(0, \text{large})$

$$\tilde{\Sigma} = (\text{large}^{-1} + I_B)^{-1} \approx I_B^{-1}$$

$$\tilde{\beta} = I_B^{-1} (\text{large}^{-1} \beta_0 + I_B \hat{\beta}_{\text{ML}})$$
$$\approx \hat{\beta}_{\text{ML}}$$

Prior 2: w.h.p. $0 < \beta < 20$

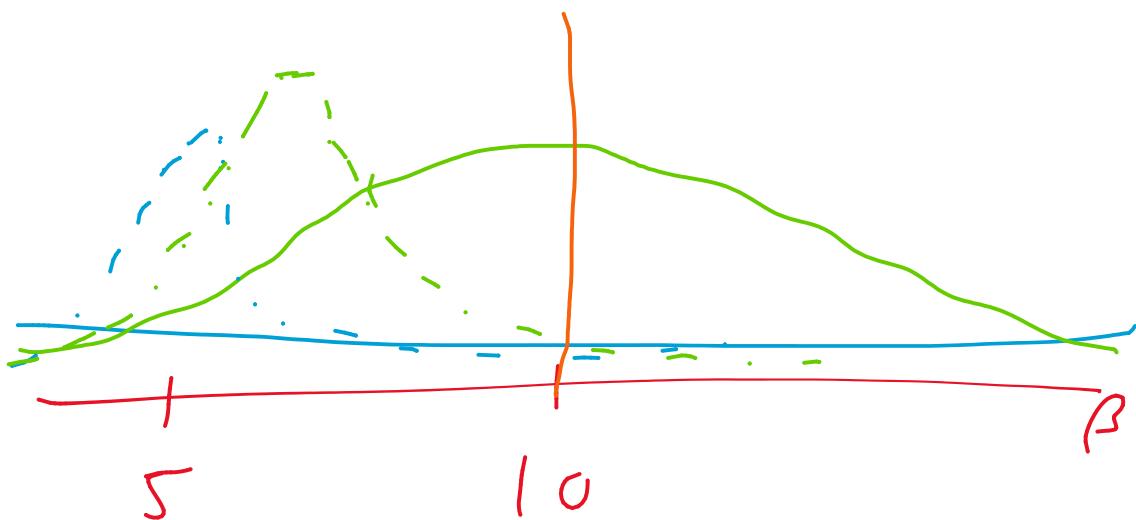
$$\beta \sim N(10, 25)$$

$$\beta \sim N(10, < \sigma)$$

\Rightarrow

$$\hat{\beta} = 5.2$$

$$\hat{\Sigma} = 25/26$$



Lecture 5/2

Wednesday, May 2, 2018 1:26 PM

Overview

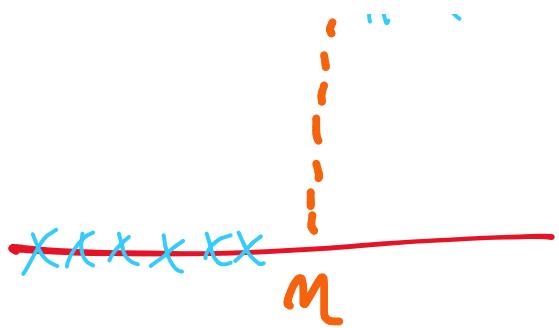
Last time: Bayesian GLM
this time: Separation &
Bayes GLM

Separation

In a logistic Regression
Separation occurs when
there is a linear
space value that

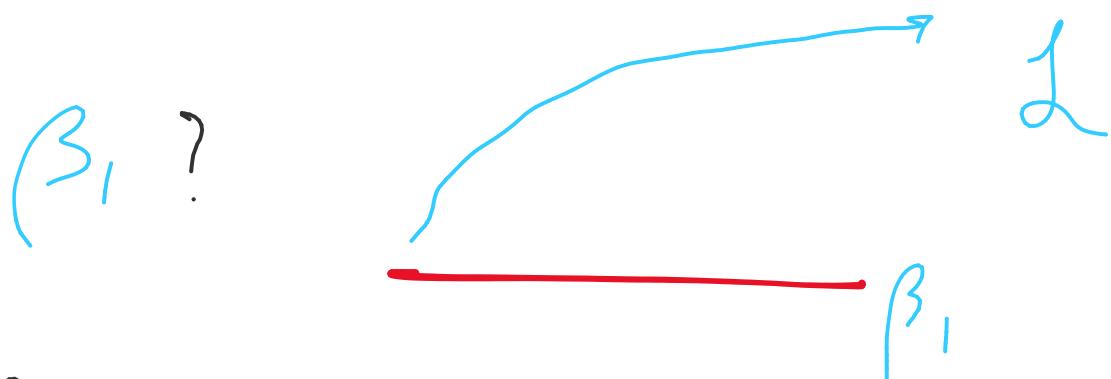
Separates the y_i





$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i$$

what's the MLE for



$$\text{So } \hat{\beta}_{1 \text{ MLE}} = \infty$$

Issues that arise.

1. IRLS will never

Converge.

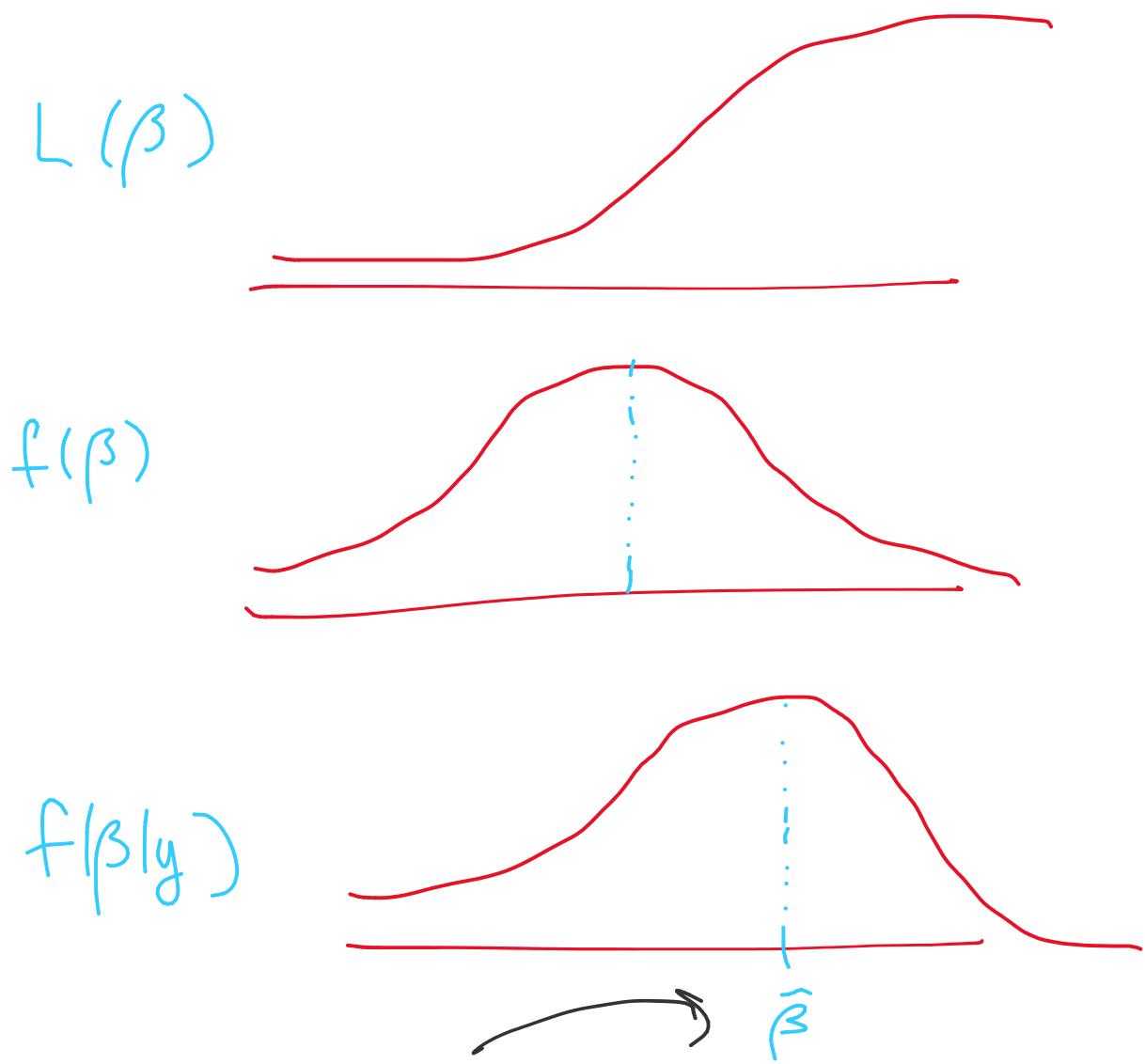
2. β values become
harder to interpret.

3. Asymptotics for I_B
will be numerically
unstable.

Solution: many approaches
including Bayesian GLM.

Priors in this context
will be used to pull
estimates from $\pm \infty$
and parameters preserved

and parameters preserve
normality

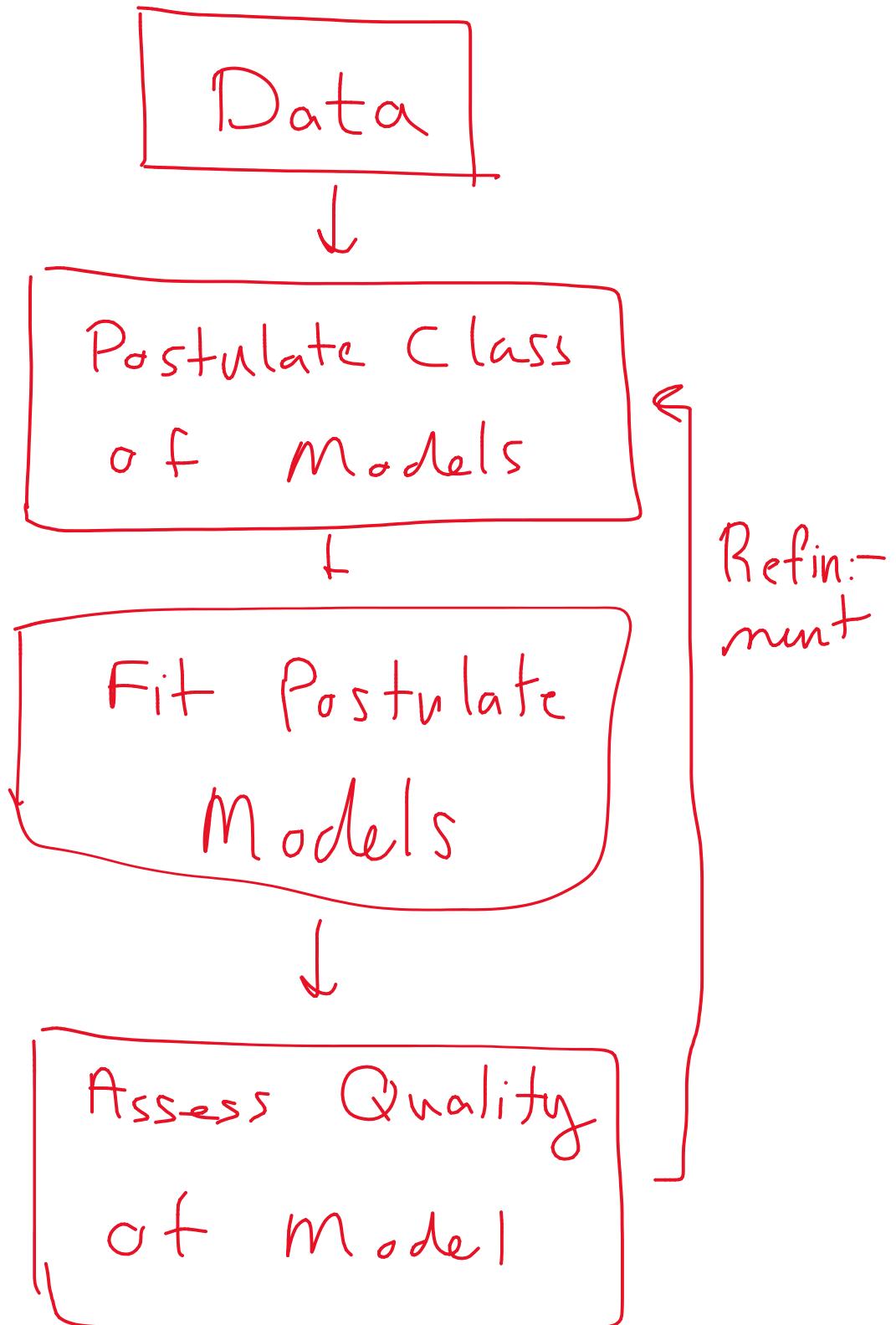


maximum a posterior (MAP)
estimate.

- Preserves estimates
- Also controls the confidence through this scheme.
- Sacrificing quality of fit

Model identification

Paradigm



Useful
model

