Optimal design of experiments

Session 7: Nonlinear models

Peter Goos

Universiteit Antwerpen

Binary data with logistic link

- example:
  - $y = 0$ or 1 (adhesion or no adhesion)
  - explanatory variable
    $x =$ time of plasma etching
  - $n = 2$ observations
- logistic regression model:

$$P(Y_i = 1) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

$$P(Y_i = 0) = \frac{1}{1 + e^{\beta_0+\beta_1 x_i}}$$

Likelihood

- likelihood function observation i

$$L_i = P(Y_i = y_i) = \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)^{1-y_i}$$
$$= \frac{e^{y_i(\beta_0+\beta_1 x_i)}}{1+e^{\beta_0+\beta_1 x_i}}$$

- log likelihood observation i

$$\ln L_i = \ln e^{y_i(\beta_0+\beta_1 x_i)} - \ln(1+e^{\beta_0+\beta_1 x_i})$$
$$= y_i(\beta_0+\beta_1 x_i) - \ln(1+e^{\beta_0+\beta_1 x_i})$$

Information matrix

- general definition observation i:

$$\mathbf{M}_i = -E\left(\frac{\partial^2 \ln L_i}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right) = E\left(\left(\frac{\partial \ln L_i}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \ln L_i}{\partial\boldsymbol{\theta}}\right)^T\right)$$

with $\boldsymbol{\theta}$ the vector of model parameters

- total information matrix

$$\mathbf{M} = \sum_{i=1}^{n} \mathbf{M}_i$$

Binary logistic regression

- $\mathbf{M}_i = -E\begin{bmatrix} \frac{\partial^2 \ln L_i}{\partial \beta_0^2} & \frac{\partial^2 \ln L_i}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ln L_i}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ln L_i}{\partial \beta_1^2} \end{bmatrix}$
- $\ln L_i = y_i(\beta_0 + \beta_1 x_i) - \ln(1 + e^{\beta_0 + \beta_1 x_i})$
- $\frac{\partial \ln L_i}{\partial \beta_0} = y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$
- $\frac{\partial \ln L_i}{\partial \beta_1} = y_i x_i - \frac{e^{\beta_0+\beta_1 x_i} x_i}{1+e^{\beta_0+\beta_1 x_i}}$

Binary logistic regression

- $$\frac{\partial^2 \ln L_i}{\partial \beta_0^2} = -\frac{\left(1+e^{\beta_0+\beta_1 x_i}\right) e^{\beta_0+\beta_1 x_i} - e^{\beta_0+\beta_1 x_i} e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} = -\frac{e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2}$$
- $$\frac{\partial^2 \ln L_i}{\partial \beta_0 \partial \beta_1} = -\frac{\left(1+e^{\beta_0+\beta_1 x_i}\right) e^{\beta_0+\beta_1 x_i} x_i - e^{\beta_0+\beta_1 x_i} e^{\beta_0+\beta_1 x_i} x_i}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} = -\frac{e^{\beta_0+\beta_1 x_i} x_i}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} = \frac{\partial^2 \ln L_i}{\partial \beta_1 \partial \beta_0}$$
- $$\frac{\partial^2 \ln L_i}{\partial \beta_1^2} = -\frac{\left(1+e^{\beta_0+\beta_1 x_i}\right) e^{\beta_0+\beta_1 x_i} x_i^2 - e^{\beta_0+\beta_1 x_i} x_i e^{\beta_0+\beta_1 x_i} x_i}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} = -\frac{e^{\beta_0+\beta_1 x_i} x_i^2}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2}$$

Information matrix

- observation i

$$\mathbf{M}_i = -E\begin{bmatrix} \frac{-e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} & \frac{-x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \\ \frac{-x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} & \frac{-x_i^2 e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \end{bmatrix}$$

- total information matrix $\mathbf{M} = \sum_{i=1}^{n} \mathbf{M}_i$
- the information matrix (and thus the amount of information) on the unknown parameters depends on the unknown parameters
- to maximize the information content of your experiment, you need a guess for β_0 and β_1

Information matrix

- observation i

$$\mathbf{M}_i = \begin{bmatrix} \frac{e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} & \frac{x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \\ \frac{x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} & \frac{x_i^2 e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \end{bmatrix}$$

- total information matrix $\mathbf{M} = \sum_{i=1}^{n} \mathbf{M}_i$
- the information matrix (and thus the amount of information) on the unknown parameters depends on the unknown parameters
- to maximize the information content of your experiment, you need a guess for β_0 and β_1

Locally optimal design

- binary.xls
- 2 examples are given:

$$\begin{cases} \text{parameterset 1}: & \beta_0 = -2 \text{ and } \beta_1 = +2 \\ \text{parameterset 2}: & \beta_0 = -2 \text{ and } \beta_1 = +3 \end{cases}$$

- set 1 leads to: $\begin{cases} x_1 = 0.228 \\ x_2 = 1.772 \end{cases}$
- set 2 leads to: $\begin{cases} x_1 = 0.152 \\ x_2 = 1.181 \end{cases}$

these designs are called locally optimal (optimal for just one set of β's)

Bayesian approach

- problem with locally optimal designs: they might not be very good for other β's
- a *Bayesian* (D-)optimal design is a design that performs well on average
- how?

$$\text{for each } \beta_i : \beta_i \sim \text{NORMAL}\,(\,a\,,\,b^2\,)$$

some density/distribution

I think β_i is around a

I'm not that sure, I might be wrong

(small b: I'm pretty sure $\leftrightarrow$ large b: unsure)

Simple example

- $\beta_0 = -2, \beta_1 = \begin{cases} 2 & \text{(50\% chance)} \\ 3 & \text{(50\% chance)} \end{cases}$ instead of normal
- construct information matrix for every set of β's
- calculate $|\mathbf{M}|$ for each set of β's: $|\mathbf{M}|_1$, $|\mathbf{M}|_2$
- what you have to maximize is the *Bayesian* D-criterion

  $0.5\,|\mathbf{M}|_1 + 0.5\,|\mathbf{M}|_2$ probability second set of β's

  probability first set of β's
- example: `Bayesian binary.xls`

  Bayesian D-optimal design: $\begin{cases} x_1 = 0.2 \\ x_2 = 1.573 \end{cases}$

Implementation normal prior distribution

- what if $\beta_i \sim$ NORMAL?
- generate "a lot" of β_i's from the normal distribution (R = number of draws)
- maximize the *Bayesian* D-criterion $\sum_{j=1}^{R} \frac{1}{R} |\mathbf{M}|_j$

  determinant for the jth set of β's you randomly drew from the normal distributions for β_i's
- this is done to approximate $\int_{\mathbb{R}^k} |\mathbf{M}|_j\, \pi(\beta)\, d\beta$

  joint probability distribution of β_i's

Implementation normal prior distribution

- usually, a Monte Carlo sample is drawn from the prior distribution
- for this to work well, you need to draw a lot of random samples
- this is computationally demanding
- solution: do not draw samples randomly but systematically
  - Halton sequences
  - Sobol sequences
  - Gaussian quadrature
- in that case, you need much fewer draws

More on *Bayesian* optimal design

- no *Bayesian* design:
  maximizing $|\mathbf{M}|$ and $\log|\mathbf{M}|$ is the same thing
- *Bayesian* design:
  maximizing $\sum_{j=1}^{R}\frac{1}{R}|\mathbf{M}|_j$ and $\sum_{j=1}^{R}\frac{1}{R}\log|\mathbf{M}|_j$ is NOT the same thing!
- see `Bayesian binary (version 2).xls`

  Bayesian D-optimal design: $\begin{cases} x_1 = 0.179 \\ x_2 = 1.419 \end{cases}$

Maximin designs

- designs that have the best possible worst case performance
- how?
  - for each set of β's, there is a locally optimal design, with determinant $|\mathbf{M}|_j^*$ for parameter set j
  - any other design will be worse than $|\mathbf{M}|_j^*$ for that set
  - how bad?

$$\left(\frac{\left|\mathbf{M}(\text{set } j)\right|}{|\mathbf{M}|_j^*}\right)^{1/p}$$

  - we compute this quantity for every set of β's
  - we focus on the smallest / worst value and maximize that value

Our example

	β	(locally) opt. design	opt. determ. $\lvert\mathbf{M}\rvert_j^*$
set 1	$\beta_0 = -2$	$x_1 = 0.228$	
	$\beta_1 = +2$	$x_2 = 1.772$	$\lvert\mathbf{M}\rvert_1^* = 0.0501$
set 2	$\beta_0 = -2$	$x_1 = 0.152$	
	$\beta_1 = +3$	$x_2 = 1.181$	$\lvert\mathbf{M}\rvert_2^* = 0.0223$

find design with information matrix $\mathbf{M}$ that maximizes

$$\min\left\{\left(\frac{|\mathbf{M}(-2,2)|}{|\mathbf{M}|_1^*}\right)^{1/2}, \left(\frac{|\mathbf{M}(-2,3)|}{|\mathbf{M}|_2^*}\right)^{1/2}\right\}$$

Our example

- `maximin binary.xls`
- maximin design $\begin{cases} x_1 = 0.18 \\ x_2 = 1.436 \end{cases}$
- this design is 94.4% efficient for both sets of β's
- this means that

$$\left(\frac{|\mathbf{M}(-2,2)|}{|\mathbf{M}|_1^*}\right)^{1/2} = \left(\frac{|\mathbf{M}(-2,3)|}{|\mathbf{M}|_2^*}\right)^{1/2} = 0.944$$

Sequential optimal design

- idea
  1. start with a small design and collect some data
  2. update your knowledge on model's parameters
  3. create a new design that uses improved knowledge
  4. repeat steps 2 and 3 as often as possible/desirable
- avoids constructing a large design based on poor prior knowledge
- this approach performs very well usually
- not always feasible

Other considerations

- the logistic regression models belong to a class of generalized linear models
- maximum likelihood estimation
- for some models, maximum likelihood theory can not be used to derive an information matrix
- this is what next slides are about

Nonlinear regression models

- general model (just one θ)

$$Y = \eta(x,\theta) + \epsilon$$
$$E(Y) = \eta(x,\theta)$$

- Taylor series expansion

$$\begin{aligned} E(Y) &= \eta(x,\theta) \\ &= \eta(x,\theta_0) + (\theta - \theta_0)\left.\frac{\partial\eta(x,\theta)}{\partial\theta}\right|_{\theta=\theta_0} + \dots \end{aligned}$$

Nonlinear regression models

- rewrite as

$$\underbrace{E(Y)-\eta(x,\theta_0)}_{\text{some response}} = \underbrace{(\theta-\theta_0)}_{\text{parameter}} \underbrace{\left.\frac{\partial\eta(x,\theta)}{\partial\theta}\right|_{\theta=\theta_0}}_{\text{function of exp. var.}}$$

$$Y^* = \beta f(x)$$

- nonlinear model with several θ's

$$Y^* = \boldsymbol{\beta}^T \mathbf{f}(\mathbf{x})$$

Information matrix

- information matrix for such a model

$$\mathbf{M} = \sum_{i=1}^{n} \mathbf{f}(\mathbf{x})\mathbf{f}^T(\mathbf{x})$$

- here

$$\mathbf{f}(\mathbf{x}) = \left.\frac{\partial\eta(\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\theta=\theta_0}$$

- so information matrix depends on unknown parameters
- thus, optimal designs depend on the unknown parameters

An example: a chemical reaction

$$A \xrightarrow{\theta_1} B \xrightarrow{\theta_2} C$$

$$Y_i = \frac{\theta_1}{\theta_1 - \theta_2}\left(e^{-\theta_2 t_i} - e^{-\theta_1 t_i}\right)$$

- Y_i = concentration of substance B
- t_i = time = explanatory variable
- $\theta_1 > \theta_2$
- e.g. $O_2 \rightarrow H_2O_2 \rightarrow H_2O$
- suppose $n = 4$, so you have to choose 4 time points t_1, t_2, t_3, t_4 at which to measure the presence of substance B

Model matrix $\mathbf{X}$

- dimension 4×2
- what should be in the columns?

$$\frac{\partial \eta}{\partial \theta_1} \text{ and } \frac{\partial \eta}{\partial \theta_2}$$

$$\text{here: } \frac{\partial Y}{\partial \theta_1} \text{ and } \frac{\partial Y}{\partial \theta_2}$$

- first column:

$$\frac{\partial Y}{\partial \theta_1} = \frac{1}{(\theta_1 - \theta_2)^2}\left((\theta_2 + \theta_1(\theta_1 - \theta_2)\, t_i)\, e^{-\theta_1 t_i} - \theta_2 e^{-\theta_2 t_i}\right)$$

- second column:

$$\frac{\partial Y}{\partial \theta_2} = \frac{1}{(\theta_1 - \theta_2)^2}\left((\theta_1 + \theta_1(\theta_1 - \theta_2)\, t_i)\, e^{-\theta_2 t_i} - \theta_1 e^{-\theta_1 t_i}\right)$$

Locally optimal design

- you need some idea about θ_1 and θ_2 before you can start
- e.g. $\theta_1 = 0.7$, $\theta_2 = 0.2$, so

$$\frac{\partial Y}{\partial \theta_1} = (0.8 + 1.4t_i)e^{-0.7t_i} - 0.8e^{-0.2t_i}$$
$$\frac{\partial Y}{\partial \theta_2} = (2.8 + 1.4t_i)e^{-0.2t_i} - 2.8e^{-0.7t_i}$$

- see nonlinear.xls