统计建模方法及其应用

曹春正

caochunzheng@nuist.edu.cn

统计建模方法及其应用

1. 参数回归模型

2. 非参数回归模型

3. 纵向数据模型

4. 数模真题分析

导语

- “All models are wrong but some are useful”——George Box.



导语

- “All models are wrong but some are useful”——George Box.
- “We buy information with assumptions”——C. H. Coombs.



导语

- “All models are wrong but some are useful”——George Box.
- “We buy information with assumptions”——C. H. Coombs.
- “Analyse problems, not data”——Peter J. Diggle.



统计建模方法及其应用

1. 参数回归模型

2. 非参数回归模型

3. 纵向数据模型

4. 数模真题分析

1. 参数回归模型

- $\boldsymbol{Y} = (y_1, \cdots, y_n)^\top$: 响应变量
  $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$: 协变量或解释变量

1. 参数回归模型

- $\boldsymbol{Y} = (y_1, \cdots, y_n)^\top$: 响应变量
  $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$: 协变量或解释变量
- $y_i = \eta_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$, iid, $i = 1, \cdots, n$.
  $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$: 线性模型
  $\eta_i = f(\boldsymbol{x}_i, \boldsymbol{\beta})$: 非线性模型

1. 参数回归模型

- $\boldsymbol{Y} = (y_1, \cdots, y_n)^\top$: 响应变量
  $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$: 协变量或解释变量
- $y_i = \eta_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$, iid, $i = 1, \cdots, n$.
  $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$: 线性模型
  $\eta_i = f(\boldsymbol{x}_i, \boldsymbol{\beta})$: 非线性模型
- 极大似然估计(对于线性模型):

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \\ \widehat{\sigma^2} &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n, \end{aligned}$$

1. 参数回归模型

- $\boldsymbol{Y} = (y_1, \cdots, y_n)^\top$: 响应变量
  $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$: 协变量或解释变量
- $y_i = \eta_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$, iid, $i = 1, \cdots, n$.
  $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$: 线性模型
  $\eta_i = f(\boldsymbol{x}_i, \boldsymbol{\beta})$: 非线性模型
- 极大似然估计(对于线性模型):

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \\ \widehat{\sigma^2} &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n, \end{aligned}$$

- Fisher信息阵:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 & \boldsymbol{0} \\ \boldsymbol{0} & n/(2\sigma^4) \end{bmatrix}.$$

例1: 针叶松数据(Atkinson and Riani, 2000, P. 292)

该数据包含70棵针叶松的测量数据，其中y表示体积(单位：立方英尺)，x_1 为树的直径(单位：英寸)，x_2为树的高度(单位：英尺).

No.	1	2	3	4	5	$\cdots$	69	70
x_1	4.6	4.4	5.0	5.1	5.1	$\cdots$	19.4	23.4
x_2	33	38	40	49	37	$\cdots$	94	104
y	2.2	2.0	3.0	4.3	3.0	$\cdots$	107.0	163.5

可选择模型为：　　下方的式子是怎么列出来的?

$$\sqrt[3]{y_i} = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i,\ i = 1, \cdots, 70,$$

或者，

$$\log(y_i) = \beta_0 + \log(x_{1i})\beta_1 + \log(x_{2i})\beta_2 + \varepsilon_i,\ i = 1, \cdots, 70.$$

matlab: [b, bint, r, rint, stats] = regress(y, X)

数据变换方法

数据变化的目的是啥?

定理：设随机变量y的方差为σ^2，均值为μ，且有函数关系$\sigma = \phi(\mu)$，$\phi(\cdot)$为正值可微函数，则经过下列变换后，

$$z = h(y) = \sigma_0 \int [1/\phi(y)]dy, \tag{1}$$

z的方差近似等于常数σ_0.

常用方差稳定化变换

变换公式($z = h(y)$)	适用情形	备注
$\sqrt{y}$	$\mathsf{Var}(\varepsilon_i) \propto \mathsf{E}(y_i)$	如误差项为Poisson分布
$\sqrt{y} + \sqrt{y+1}$	同上	某些y_i等于或近似等于0
$y^{1\mp k}$	$\mathsf{Var}(\varepsilon_i) \propto [\mathsf{E}(y_i)]^{1\pm 2k}$	k为大于1的整数($y_i > 0$)
$\log(y)$	$\mathsf{Var}(\varepsilon_i) \propto \mathsf{E}^2(y_i)$	$y_i > 0$
$\log(y+1)$	同上	$y_i \geq 0$
$(y+1)^{-1}$	$\mathsf{Var}(\varepsilon_i) \propto \mathsf{E}^4(y_i)$	某些y_i为0
$arcsin(\sqrt{y})$	$\mathsf{Var}(\varepsilon_i) \propto \mathsf{E}(y_i)(1-\mathsf{E}(y_i))$	如二项分布($0 \leq y_i \leq 1$)
$(1-y)^{1/2} - \frac{1}{3}(1-y)^{3/2}$	$\mathsf{Var}(\varepsilon_i) \propto (1-\mathsf{E}(y_i))/\mathsf{E}(y_i)$	如负二项分布($0 \leq y_i \leq 1$)
$\log(\frac{1+y}{1-y})$	$\mathsf{Var}(\varepsilon_i) \propto [1-\mathsf{E}^2(y_i)]^{-4}$	$-1 \leq y_i \leq 1$

Box-Cox变换

Tukey, 1957:

$$y(\lambda) = \begin{cases} y^{\lambda} & \lambda \neq 0; \\ \log(y) & \lambda = 0. \end{cases} \tag{2}$$

Box and Cox, 1964:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0; \\ \log(y) & \lambda = 0. \end{cases} \tag{3}$$

例2: 欧洲野兔数据(Dudzinski and Mykytowycz,1961)

No.	1	2	3	4	5	$\cdots$	70	71
x	15	15	15	18	28	$\cdots$	768	860
y	21.66	22.75	22.30	31.25	44.79	$\cdots$	232.12	246.70

其中y为在澳大利亚的欧洲野兔干燥眼球重量(单位:毫克),x为野兔相应的年龄(单位: 天). 其模型为

$$\log(y_i) = \beta_1 - \beta_2(x_i + \beta_3)^{-1} + \varepsilon_i, i = 1, \cdots, 71.$$

上方的式子在表格中没找到,
所以上方的式子是怎么列出来的.

- 极大似然估计:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (\dot{\boldsymbol{F}}^\top \dot{\boldsymbol{F}})^{-1} \dot{\boldsymbol{F}}^\top \boldsymbol{\varepsilon}|_{\boldsymbol{\theta}^{(k)}},$$

$$\sigma^{2(k+1)} = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}/n|_{\boldsymbol{\theta}^{(k)}},$$

这里, $\dot{\boldsymbol{F}} = (\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial \beta_j)_{n \times p}$.

例2: 欧洲野兔数据(Dudzinski and Mykytowycz,1961)

No.	1	2	3	4	5	$\cdots$	70	71
x	15	15	15	18	28	$\cdots$	768	860
y	21.66	22.75	22.30	31.25	44.79	$\cdots$	232.12	246.70

其中y为在澳大利亚的欧洲野兔干燥眼球重量(单位:毫克),x为野兔相应的年龄(单位: 天). 其模型为

$$\log(y_i) = \beta_1 - \beta_2(x_i + \beta_3)^{-1} + \varepsilon_i, i = 1, \cdots, 71.$$

- 极大似然估计:

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} =& \boldsymbol{\beta}^{(k)} + (\dot{\boldsymbol{F}}^\top \dot{\boldsymbol{F}})^{-1}\dot{\boldsymbol{F}}^\top \boldsymbol{\varepsilon}|_{\boldsymbol{\theta}^{(k)}},\\ \sigma^{2(k+1)} =& \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}/n|_{\boldsymbol{\theta}^{(k)}},\end{aligned}$$

这里, $\dot{\boldsymbol{F}} = (\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial\beta_j)_{n\times p}$.

- Fisher信息阵:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{bmatrix} \dot{\boldsymbol{F}}^\top \dot{\boldsymbol{F}}/\sigma^2 & \mathbf{0} \\ \mathbf{0} & n/(2\sigma^4) \end{bmatrix}.$$

例3(分类数据模型): 降雨数据

年份	x_1	x_2	x_3	x_4	y
1951	0.58	82.0	44.0	40.6	1
1952	0.40	83.0	18.0	43.0	3
1953	0.55	85.0	36.0	30.7	3
...	...	...	...	...	...
1973	0.53	83.0	23.0	61.3	2
1974	0.48	84.0	19.0	23.2	3
1975	0.30	85.0	27.0	17.5	3

北京市25年有关降雨资料, x_1, x_2, x_3, x_4是4个预报因子，y表示降雨情况：$y=1$表示偏少, $y=2$表示正常, $y=3$表示偏多.

- 适用模型：广义线性模型(Logistic模型或Probit模型)

$$\text{Logit}(\pi_i)=\log(\frac{\pi_i}{1-\pi_i})=\boldsymbol{x}_i^\top\boldsymbol{\beta}, i=1,\cdots,n.$$

例3(分类数据模型): 降雨数据

年份	x_1	x_2	x_3	x_4	y
1951	0.58	82.0	44.0	40.6	1
1952	0.40	83.0	18.0	43.0	3
1953	0.55	85.0	36.0	30.7	3
...	...	...	...	...	...
1973	0.53	83.0	23.0	61.3	2
1974	0.48	84.0	19.0	23.2	3
1975	0.30	85.0	27.0	17.5	3

北京市25年有关降雨资料, x_1, x_2, x_3, x_4是4个预报因子，y表示降雨情况：$y=1$表示偏少, $y=2$表示正常, $y=3$表示偏多.

- 适用模型：广义线性模型(Logistic模型或Probit模型)

$$\text{Logit}(\pi_i)=\log(\frac{\pi_i}{1-\pi_i})=\boldsymbol{x}_i^\top\boldsymbol{\beta}, i=1,\cdots,n.$$

- 极大似然：$L(\boldsymbol{\beta})=\prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i}$.

- 多项Logistic回归(无序)：

$$\pi_{ik} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}, \ i = 1, \ldots, n, j = 1, \ldots, K.$$

- 多项Logistic回归(无序)：

$$\pi_{ik} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}, \ i = 1, \ldots, n, j = 1, \ldots, K.$$

- 多项Logistic回归(有序)：

$$\log(\frac{\sum_{j=1}^{k} \pi_{ij}}{1 - \sum_{j=1}^{k} \pi_{ij}}) = \beta_0^{(k)} + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \ i = 1, \ldots, n, j = 1, \ldots, K-1.$$

- 多项Logistic回归(无序)：

$$\pi_{ik} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}, \ i = 1, \ldots, n, j = 1, \ldots, K.$$

- 多项Logistic回归(有序)：

$$\log(\frac{\sum_{j=1}^{k} \pi_{ij}}{1 - \sum_{j=1}^{k} \pi_{ij}}) = \beta_0^{(k)} + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \ i = 1, \ldots, n, j = 1, \ldots, K-1.$$

- matlab: glmfit

统计建模方法及其应用

1. 参数回归模型

2. 非参数回归模型

3. 纵向数据模型

4. 数模真题分析

2. 非参数回归模型

模型一般形式:

$$y_i = f(x_i) + \varepsilon_i, \mathsf{E}(\varepsilon_i) = 0, \ i = 1, \cdots, n. \tag{4}$$

方法：

- 核回归

2. 非参数回归模型

模型一般形式:

$$y_i = f(x_i) + \varepsilon_i, \mathsf{E}(\varepsilon_i) = 0, \ i = 1, \cdots, n. \tag{4}$$

方法：

- 核回归
- 局部多项式回归

2. 非参数回归模型

模型一般形式:

$$y_i = f(x_i) + \varepsilon_i, \mathsf{E}(\varepsilon_i) = 0,\ i = 1, \cdots, n. \tag{4}$$

方法：

- 核回归
- 局部多项式回归
- 样条

2. 非参数回归模型

模型一般形式:

$$y_i = f(x_i) + \varepsilon_i, \mathsf{E}(\varepsilon_i) = 0, \ i = 1, \cdots, n. \tag{4}$$

方法：

- 核回归
- 局部多项式回归
- 样条
- .....

核估计

定义1：线性光滑器(linear smoother)

如果对于每个点x，存在一个向量$\boldsymbol{l}(x) = (l_1(x), \cdots, l_n(x))^\top$，使得$r$的一个估计具备形式

$$\widehat{f}_n(x) = \sum_{i=1}^{n} l_i(x) y_i, \tag{5}$$

则称估计$\widehat{r}_n$为一个线性光滑器。

定义2：Nadaraya-Watson核估计

若线性光滑器中权重$l_i(x)$有如下形式:

$$l_i(x) = K(\frac{x - x_i}{h}) \Big/ \sum_{j=1}^{n} K(\frac{x - x_j}{h}), \tag{6}$$

则称估计$\widehat{f}_n$为Nadaraya-Watson核估计, 其中$K(\cdot)$为核函数, 参数h为带宽(bandwidth).

(1) 核函数的选取：

- boxcar核：$K(u) = \frac{1}{2} I(u)$.

这里$I(\cdot)$为示性函数，满足

$$I(u) = \begin{cases} 1, & |u| \leqslant 1, \\ 0, & |u| > 1. \end{cases}$$

(1) 核函数的选取：

- boxcar核：$K(u) = \frac{1}{2} I(u)$.
- Gaussian核：$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$.

这里$I(\cdot)$为示性函数，满足

$$I(u) = \begin{cases} 1, & |u| \leqslant 1, \\ 0, & |u| > 1. \end{cases}$$

(1) 核函数的选取：

- boxcar核：$K(u) = \frac{1}{2}I(u)$.
- Gaussian核：$K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$.
- Epanechnikov核: $K(u) = \frac{3}{4}(1-u^2)I(u)$.

这里$I(\cdot)$为示性函数，满足

$$I(u) = \begin{cases} 1, & |u| \leqslant 1, \\ 0, & |u| > 1. \end{cases}$$

(1) 核函数的选取：

- boxcar核：$K(u) = \frac{1}{2}I(u)$.
- Gaussian核：$K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$.
- Epanechnikov核: $K(u) = \frac{3}{4}(1-u^2)I(u)$.
- tricube核：$K(u) = \frac{70}{81}(1-|u|^3)^3 I(u)$.

这里$I(\cdot)$为示性函数，满足

$$I(u) = \begin{cases} 1, & |u| \leqslant 1, \\ 0, & |u| > 1. \end{cases}$$

(2) 带宽h的选择：

- 缺一交叉验证(leave-one-out cross-validation):

$$CV(h)=\frac{1}{n}\sum_{i=1}^{n}[y_i-\widehat{f}_{(-i)}(x_i)]^2=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i-\widehat{f}_n(x_i)}{1-L_{ii}}\right]^2, \quad (7)$$

这里$\widehat{f}_{(-i)}$指未用数据点(x_i,y_i)时所得到的估计，L_{ii}为光滑矩阵L的第i个对角元，其中$L=(\boldsymbol{l}(x_1),\cdots,\boldsymbol{l}(x_n))^\top$.

$$\widehat{f}_n(x)=\sum_{i=1}^{n}l_i(x)y_i,$$

(2) 带宽h的选择：

- 缺一交叉验证(leave-one-out cross-validation):

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \widehat{f}_{(-i)}(x_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \widehat{f}_n(x_i)}{1 - L_{ii}}\right]^2, \quad (7)$$

这里$\widehat{f}_{(-i)}$指未用数据点(x_i, y_i)时所得到的估计，L_{ii}为光滑矩阵L的第i个对角元，其中$L = (\boldsymbol{l}(x_1), \cdots, \boldsymbol{l}(x_n))^\top$.

- 广义交叉验证(generalized cross-validation):

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \widehat{f}_n(x_i)}{1 - \nu/n}\right]^2, \quad (8)$$

其中，$\nu = \mathrm{tr}(L)$称为有效自由度.

局部多项式估计

估计思想：

- 最小化$\sum_{i=1}^{n}(y_i - a)^2 \Longrightarrow \widehat{f}_n(x) = \overline{y}$. (样本均值)

局部多项式估计

估计思想：

- 最小化$\sum_{i=1}^{n}(y_i-a)^2 \Longrightarrow \widehat{f}_n(x)=\overline{y}$. (样本均值)
- 最小化$\sum_{i=1}^{n} w_i(x)(y_i-a)^2 \Longrightarrow$
  $\widehat{f}_n(x)=\sum_{i=1}^{n} w_i(x)y_i \Big/ \sum_{i=1}^{n} w_i(x)$. (核估计)

权重的 取值方法, 有啥注意点, 有啥技巧吗?

局部多项式估计

估计思想：

- 最小化$\sum_{i=1}^{n}(y_i-a)^2 \Longrightarrow \widehat{f}_n(x)=\overline{y}$. (样本均值)
- 最小化$\sum_{i=1}^{n} w_i(x)(y_i-a)^2 \Longrightarrow$
  $\widehat{f}_n(x)=\sum_{i=1}^{n} w_i(x)y_i \Big/ \sum_{i=1}^{n} w_i(x)$. (核估计)
- 最小化$\sum_{i=1}^{n} w_i(x)(y_i-P_x(x_i;\boldsymbol{a}))^2$,其中多项式

$$P_x(u;\boldsymbol{a})=a_0+a_1(u-x)+\cdots+\frac{a_p}{p!}(u-x)^p.$$

$\Longrightarrow \widehat{f}_n(x)=P_x(x;\widehat{\boldsymbol{a}})=\widehat{a}_0(x)$.
注：$p=0$即为核估计，$p=1$称为局部线性估计.

定理：局部多项式估计可表示为

$$\widehat{f}_n(x) = \sum_{i=1}^{n} l_i(x) y_i, \tag{9}$$

其中$\boldsymbol{l} = \boldsymbol{e}_1^\top (\boldsymbol{X}_x^\top \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} \boldsymbol{X}_x^\top \boldsymbol{W}_x$, 这里$\boldsymbol{e}_1 = (1, 0, \cdots, 0)^\top$为$p+1$维列向量,

$$\boldsymbol{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & \frac{(x_1-x)^p}{p!} \\ 1 & x_2 - x & \cdots & \frac{(x_2-x)^p}{p!} \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & \frac{(x_n-x)^p}{p!} \end{bmatrix},$$

而$\boldsymbol{W}_x = \mathsf{diag}\{w_1(x), \cdots, w_n(x)\}$.

注: 局部多项式的带宽选择一般仍采用交叉验证法.

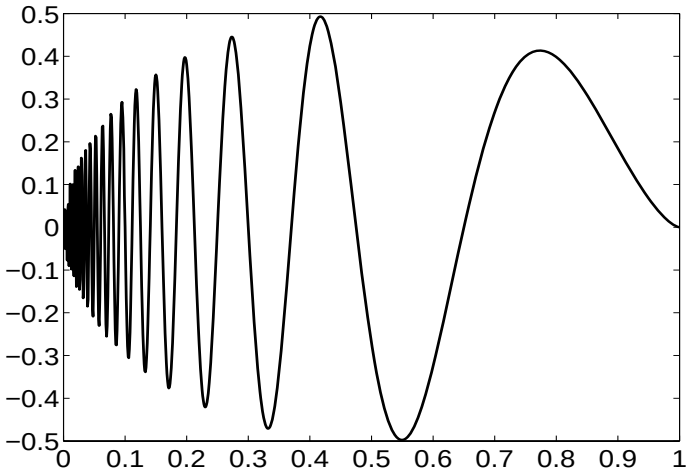
例4：非参数模拟实验
数据产生自

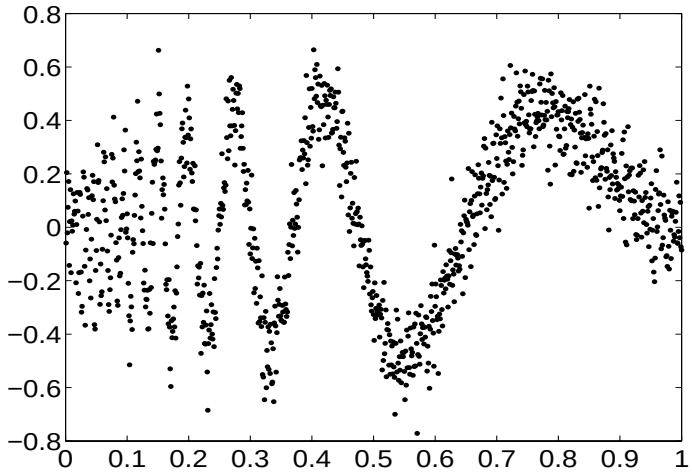
$$y_i = f(i/n) + \sigma\varepsilon_i,\ i = 1, \cdots, n,$$

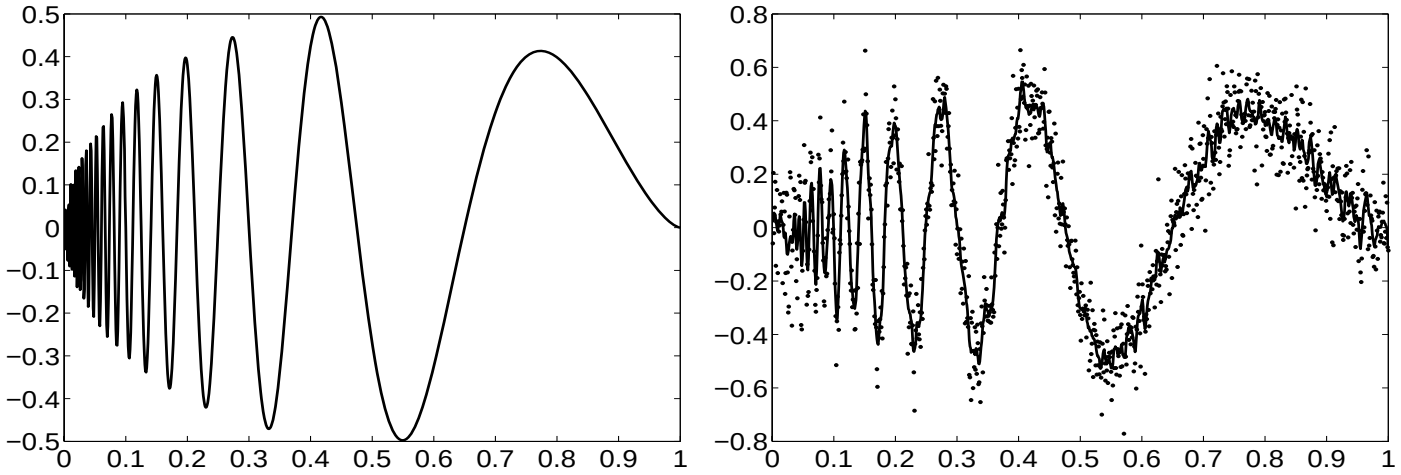
其中, $n = 1000$, $\sigma = 0.1$, $\varepsilon_i \sim N(0,1)$, 而

$$f(x) = \sqrt{x(1-x)}\sin(\frac{2.1\pi}{x+0.05}),\ 0 \leqslant x \leqslant 1,$$

称为Doppler函数.







由局部线性估计得到的估计效果，其中核函数采用Epanechnikov核，带宽选择采用缺一交叉验证($h = 0.004$).

统计建模方法及其应用

1. 参数回归模型
2. 非参数回归模型
3. 纵向数据模型
4. 数模真题分析

3. 纵向数据模型

纵向数据分析是统计学的热点, 其主要研究特征是探索不同个体在时间或空间上的重复测量数据的统计性质. 如何建立“较好的模型”来拟合相应的纵向数据是后续统计分析的前提.

一般来说，可供选择的模型有：

- 参数模型

3. 纵向数据模型

纵向数据分析是统计学的热点, 其主要研究特征是探索不同个体在时间或空间上的重复测量数据的统计性质. 如何建立“较好的模型”来拟合相应的纵向数据是后续统计分析的前提.

一般来说，可供选择的模型有：

- 参数模型
- 非参数模型

3. 纵向数据模型

纵向数据分析是统计学的热点, 其主要研究特征是探索不同个体在时间或空间上的重复测量数据的统计性质. 如何建立“较好的模型”来拟合相应的纵向数据是后续统计分析的前提.

一般来说，可供选择的模型有：

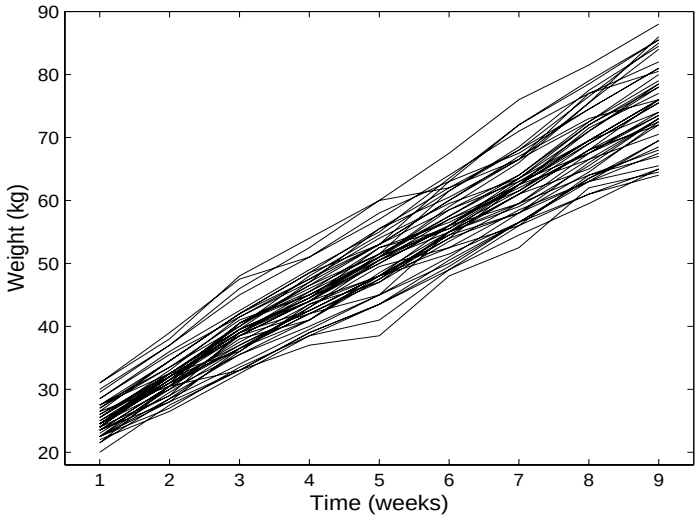
- 参数模型
- 非参数模型
- 半参数模型

例5：猪数据（Provided by Dr. Philip McCloud of Monash University at Melbourne）

	week								
No.	1	2	3	4	5	6	7	8	9
1	24.0	32.0	39.0	42.5	48.0	54.5	61.0	65.0	72.0
2	22.5	30.5	40.5	45.0	51.0	58.5	64.0	72.0	78.0
3	22.5	28.0	36.5	41.0	47.5	55.0	61.0	68.0	76.0
4	24.0	31.5	39.5	44.5	51.0	56.0	59.5	64.0	67.0
5	24.5	31.5	37.0	42.5	48.0	54.0	58.0	63.0	65.5
6	23.0	30.0	35.5	41.0	48.0	51.5	56.5	63.5	69.5
7	22.5	28.5	36.0	43.5	47.0	53.5	59.5	67.5	73.5
⋮									
47	29.5	37.0	46.0	52.5	60.0	67.5	76.0	81.5	88.0
48	28.5	36.0	42.5	49.0	55.0	63.5	72.0	78.5	85.5

注: 48头猪连续9个星期的体重测量值

例5的模型讨论:



48头猪连续9周的体重增长图

特点?

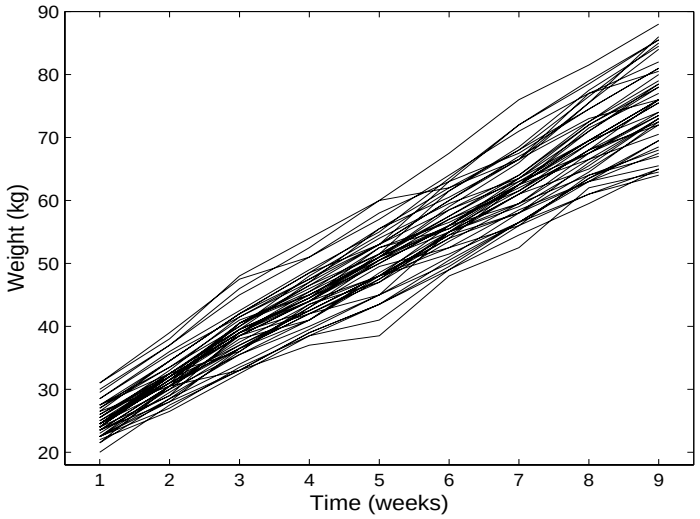
- 线性递增

$$y_{ij} = \alpha + \beta x_j + \varepsilon_{ij},$$

$$x_j = j, \quad \varepsilon_{ij} \sim N(0, \sigma^2),$$

$$i = 1, \cdots, 48, \quad j = 1, \cdots, 9.$$

例5的模型讨论:



48头猪连续9周的体重增长图

特点?

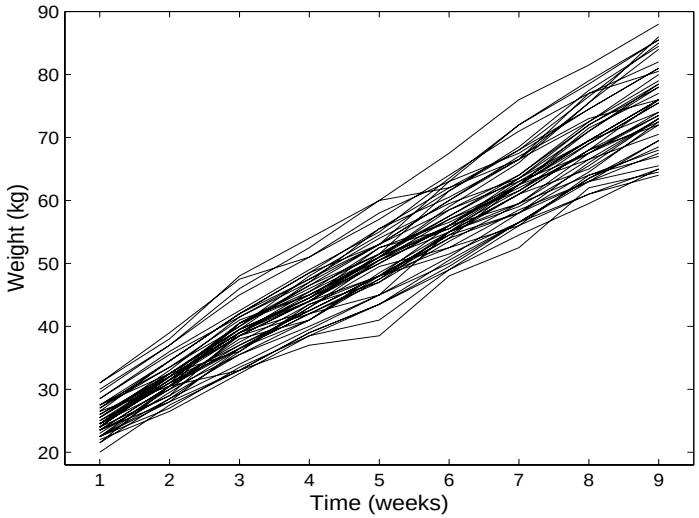
- 线性递增

$$y_{ij} = \alpha + \beta x_j + \varepsilon_{ij},$$
$$x_j = j, \quad \varepsilon_{ij} \sim N(0, \sigma^2),$$
$$i = 1, \cdots, 48, \quad j = 1, \cdots, 9.$$

- 初始体重的差异

$$y_{ij} = \alpha + \beta x_j + U_i + \varepsilon_{ij},$$
$$U_i \sim N(0, \nu^2).$$

例5的模型讨论:



48头猪连续9周的体重增长图

特点?

- 线性递增

$$y_{ij} = \alpha + \beta x_j + \varepsilon_{ij},$$
$$x_j = j, \quad \varepsilon_{ij} \sim N(0, \sigma^2),$$
$$i = 1, \cdots, 48, \quad j = 1, \cdots, 9.$$

- 初始体重的差异

$$y_{ij} = \alpha + \beta x_j + U_i + \varepsilon_{ij},$$
$$U_i \sim N(0, \nu^2).$$

- 斜率的变化

$$y_{ij} = \alpha + \beta x_j + U_i + W_i x_j + \varepsilon_{ij},$$
$$W_i \sim N(0, \tau^2).$$

最终模型:

$$
\begin{aligned}
&y_{ij} = \alpha + \beta x_j + U_i + W_i x_j + \varepsilon_{ij}, \\
&\varepsilon_{ij} \sim N(0, \sigma^2), \\
&U_i \sim N(0, \nu^2), \quad W_i \sim N(0, \tau^2).
\end{aligned} \tag{10}
$$

向量形式:

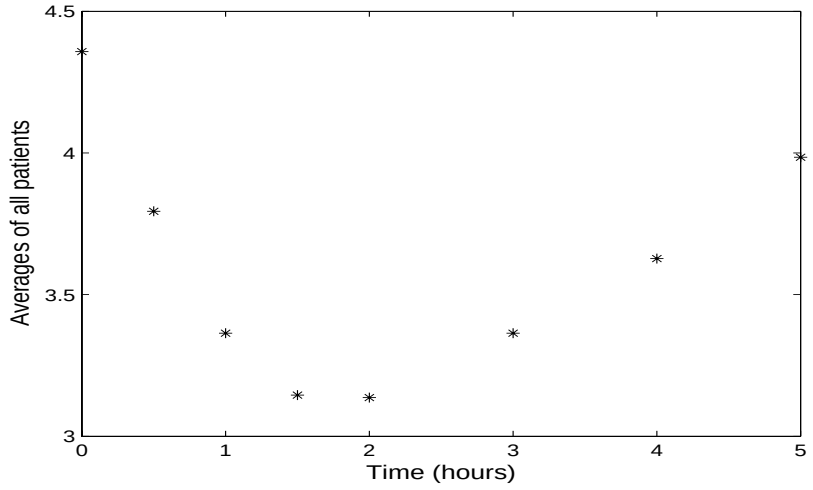
$$
\begin{aligned}
&\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, \\
&\boldsymbol{y}_i = (y_{i,1}, \cdots, y_{i,9})^\top, \boldsymbol{\beta} = (\alpha, \beta)^\top, \\
&\boldsymbol{X}_i = \boldsymbol{Z}_i = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{bmatrix}^\top, \\
&\mathbf{b}_i = (U_i, W_i)^\top \sim N_2(\mathbf{0}, \boldsymbol{D}), \boldsymbol{D} = \mathsf{diag}\{\nu^2, \tau^2\}, \\
&\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{i9})^\top \sim N_9(\mathbf{0}, \sigma^2 \boldsymbol{I}_9).
\end{aligned}
$$

例6：葡萄糖数据（Reported by Zerbe, 1979）

	hour (Control)							
No.	0	0.5	1	1.5	2	3	4	5
1	4.3	3.3	3.0	2.6	2.2	2.5	3.4	4.4
2	3.7	2.6	2.6	1.9	2.9	3.2	3.1	3.9
⋮								
13	4.7	3.1	3.2	3.3	3.2	4.2	3.7	4.3
	hour (Obese)							
No.	0	0.5	1	1.5	2	3	4	5
14	4.3	3.3	3.0	2.6	2.2	2.5	2.4	3.4
15	5.0	4.9	4.1	3.7	3.7	4.1	4.7	4.9
⋮								
33	4.6	4.4	3.8	3.8	3.8	3.6	3.8	3.8

注：20个肥胖病人和13个对照者的葡萄糖耐受试验, 数据（血浆中无机磷酸盐）从服用标准剂量的葡萄糖后0, 0.5, 1, 1.5, 2, 3, 4和5小时的试验者的血样里取得，目的是研究对照组和肥胖病人组是否有显著差异.

例6的模型讨论：



全体病人无机磷平均测量值图

抛物线模型:

下方的Z_i 是个啥?

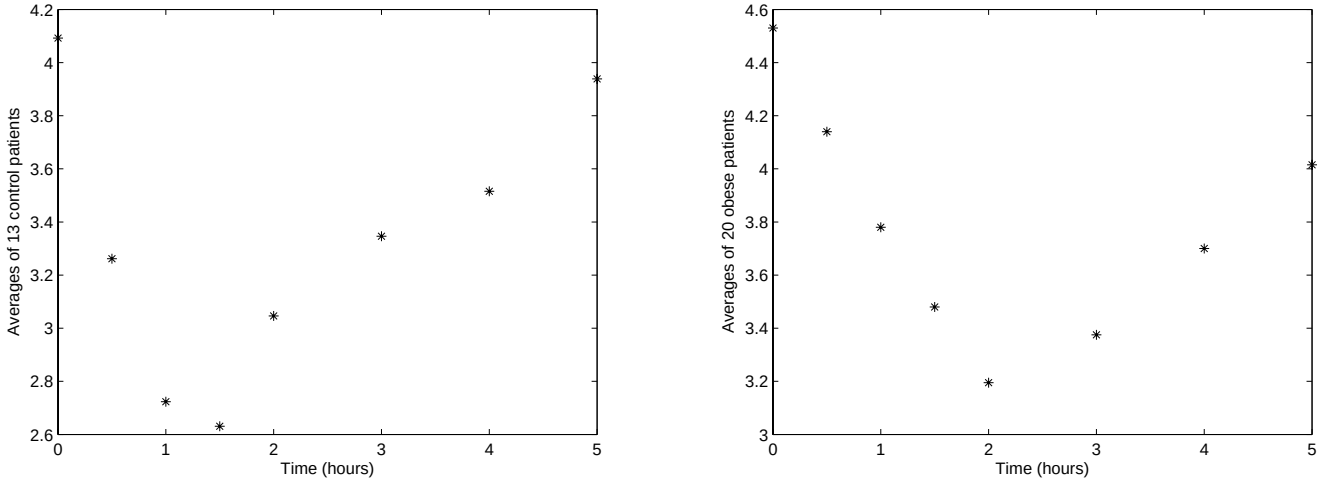
$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathsf{b}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \cdots, 33,$$

$$\boldsymbol{y}_i = (y_{i1}, \cdots, y_{i8})^\top,\ \boldsymbol{X}_i = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 1 & 1.5 & 2 & 3 & 4 & 5 \\ 0 & 0.25 & 1 & 2.25 & 4 & 9 & 16 & 25 \end{bmatrix}^\top,$$

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top, \boldsymbol{Z}_i = \mathbf{1}_8,$$

$$\mathsf{b}_i \sim N(0, \tau^2), \boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{i8})^\top \sim N_8(\mathbf{0}, \sigma^2\boldsymbol{I}_8).$$

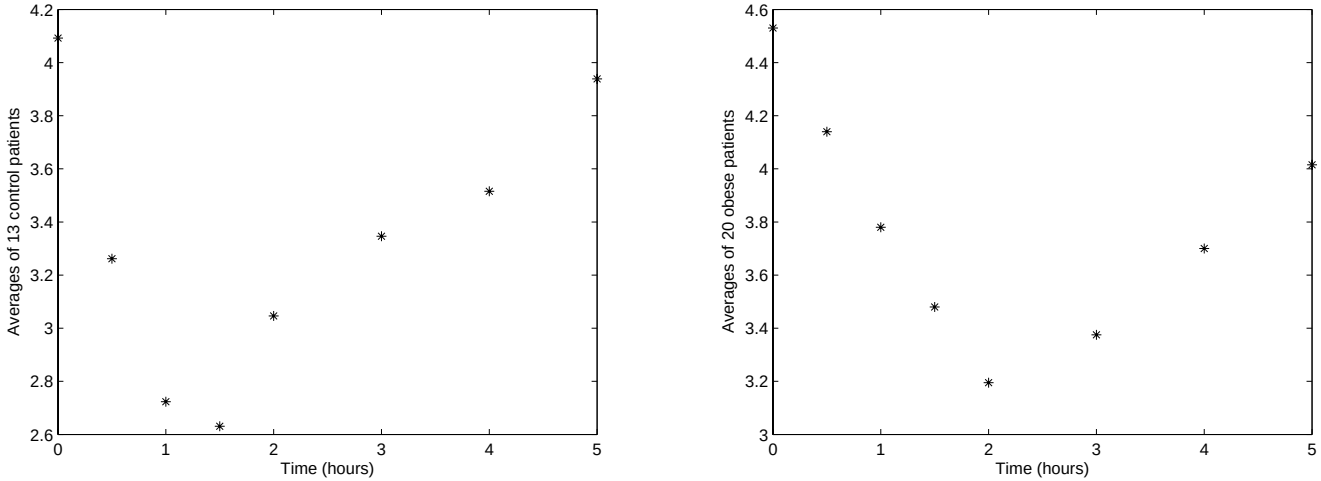
考虑组别差异:



分组无机磷平均测量值图（左为控制组，右为肥胖组）

- 分别用抛物线模型拟合.

考虑组别差异:



分组无机磷平均测量值图（左为控制组，右为肥胖组）

- 分别用抛物线模型拟合.
- 或者？

可用分段线性模型.

- 以肥胖组为例:

$$
\begin{aligned}
&\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathsf{b}_i + \boldsymbol{\varepsilon}_i,\ i = 14, \cdots, 33,\\
&\boldsymbol{y}_i = (y_{i1}, \cdots, y_{i8})^\top,\\
&\boldsymbol{X}_i = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 1 & 1.5 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 \end{bmatrix}^\top,\\
&\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top, \boldsymbol{Z}_i = \mathbf{1}_8,\\
&\mathsf{b}_i \sim N(0, \tau^2), \boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{i8})^\top \sim N_8(\mathbf{0}, \sigma^2\boldsymbol{I}_8).
\end{aligned}
$$

这个b_i 到底怎么取，tao取什么，sigma取什么

线性混合效应模型总结:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathsf{b}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \cdots, n,\ j = 1, \cdots, m_i.$$

后续工作:

- 参数估计: 极大似然估计、约束极大似然估计等方法.

线性混合效应模型总结:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathsf{b}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \cdots, n,\ j = 1, \cdots, m_i.$$

后续工作:

- 参数估计: 极大似然估计、约束极大似然估计等方法.
- 模型检验: 参数、结构.....

线性混合效应模型总结:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathsf{b}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \cdots, n,\ j = 1, \cdots, m_i.$$

后续工作:

- 参数估计: 极大似然估计、约束极大似然估计等方法.
- 模型检验: 参数、结构.....
- 预测分析: 基于经验贝叶斯方法.

$$\widehat{\boldsymbol{y}}_i = \sigma^2\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\boldsymbol{\beta} + (\boldsymbol{I}_{m_i} - \sigma^2\boldsymbol{\Sigma}_i^{-1})\boldsymbol{y}_i\Big|_{\widehat{\boldsymbol{\theta}}},$$

其中, $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\top + \sigma^2\boldsymbol{I}_{m_i}$, $\boldsymbol{\theta}$为模型全参数.

线性混合效应模型总结:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathsf{b}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \cdots, n,\ j = 1, \cdots, m_i.$$

后续工作:

- 参数估计: 极大似然估计、约束极大似然估计等方法.
- 模型检验: 参数、结构.....
- 预测分析: 基于经验贝叶斯方法.

$$\widehat{\boldsymbol{y}}_i = \sigma^2\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\boldsymbol{\beta} + (\boldsymbol{I}_{m_i} - \sigma^2\boldsymbol{\Sigma}_i^{-1})\boldsymbol{y}_i\Big|_{\widehat{\boldsymbol{\theta}}},$$

其中, $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\top + \sigma^2\boldsymbol{I}_{m_i}$, $\boldsymbol{\theta}$为模型全参数.

- .....

广义线性混合效应模型（Generalized Linear Mixed-Effects Models）

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim \text{Distr}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}/w_i),$$
$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n.$$

数据类型：Normal, Gamma, InverseGaussian, Poisson, Binomial, ...

联系函数：

- comploglog: $g(\mu) = \log(\log(1 - \mu))$;

matlab: fitglme.m

广义线性混合效应模型（Generalized Linear Mixed-Effects Models）

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim \text{Distr}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}/w_i),$$
$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n.$$

数据类型：Normal, Gamma, InverseGaussian, Poisson, Binomial, ...

联系函数：

- comploglog: $g(\mu) = \log(\log(1-\mu))$;
- log: $g(\mu) = \log(\mu)$;

matlab: fitglme.m

广义线性混合效应模型（Generalized Linear Mixed-Effects Models）

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim \text{Distr}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}/w_i),$$
$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n.$$

数据类型：Normal, Gamma, InverseGaussian, Poisson, Binomial, ...

联系函数：

- comploglog: $g(\mu) = \log(\log(1-\mu))$;
- log: $g(\mu) = \log(\mu)$;
- logit: $g(\mu) = \log(\mu/(1-\mu))$;

matlab: fitglme.m

广义线性混合效应模型（Generalized Linear Mixed-Effects Models）

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim \text{Distr}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}/w_i),$$
$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n.$$

数据类型：Normal, Gamma, InverseGaussian, Poisson, Binomial, ...

联系函数：

- comploglog: $g(\mu) = \log(\log(1-\mu))$;
- log: $g(\mu) = \log(\mu)$;
- logit: $g(\mu) = \log(\mu/(1-\mu))$;
- probit: $g(\mu) = \Phi^{-1}(\mu)$;

matlab: fitglme.m

统计建模方法及其应用

1. 参数回归模型

2. 非参数回归模型

3. 纵向数据模型

4. 数模真题分析

4. 数模真题分析

艾滋病是当前人类社会最严重的瘟疫之一, 从1981年发现以来的20多年间, 它已经吞噬了近3000万人的生命.

艾滋病的医学全名为"获得性免疫缺损综合症", 英文简称AIDS, 它是由艾滋病毒（医学全名为"人体免疫缺损病毒", 英文简称HIV）引起的. 这种病毒破坏人的免疫系统, 使人体丧失抵抗各种疾病的能力, 从而严重危害人的生命. 人类免疫系统的CD4细胞在抵御HIV的入侵中起着重要作用, 当CD4被HIV感染而裂解时, 其数量会急剧减少, HIV将迅速增加, 导致AIDS发作.

艾滋病治疗的目的, 是尽量减少人体内HIV的数量, 同时产生更多的CD4, 至少要有效地降低CD4减少的速度, 以提高人体免疫能力.

迄今为止人类还没有找到能根治AIDS的疗法, 目前的一些AIDS疗法不仅对人体有副作用, 而且成本也很高. 许多国家和医疗组织都在积极试验、寻找更好的AIDS疗法.

现在得到了美国艾滋病医疗试验机构ACTG公布的两组数据:

- ACTG320（见附件1）是同时服用zidovudine（齐多夫定），lamivudine（拉美夫定）和indinavir（茚地那韦）3种药物的300多名病人每隔几周测试的CD4和HIV的浓度（每毫升血液里的数量）;

现在得到了美国艾滋病医疗试验机构ACTG公布的两组数据:

- ACTG320（见附件1）是同时服用zidovudine（齐多夫定），lamivudine（拉美夫定）和indinavir（茚地那韦）3种药物的300多名病人每隔几周测试的CD4和HIV的浓度（每毫升血液里的数量）;
- 193A（见附件2）是将1300多名病人随机地分为4组, 每组按下述4种疗法中的一种服药, 大约每隔8周测试的CD4浓度（这组数据缺HIV浓度, 它的测试成本很高）. 4种疗法的日用药分别为: 600mg zidovudine或400mg didanosine（去羟基苷）, 这两种药按月轮换使用: 600mg zidovudine加2.25 mg zalcitabine（扎西他滨）; 600 mg zidovudine加400 mg didanosine; 600 mg zidovudine加400 mg didanosine, 再加400 mg nevirapine（奈韦拉平）.

附件1: ACTG320数据（同时服用3种药物(zidovudine, lamivudine, indinavir) 的300多名病人每隔几周测试的CD4和HIV的浓度）:

PtID	CD4Date	CD4Count	RNADate	VLoad
23424	0	178	0	5.5
23424	4	228	4	3.9
23424	8	126	8	4.7
23424	25	171	25	4
23424	40	99	40	5
23425	0	14	0	5.3
23425	4	62	4	2.4
23425	9	110	9	3.7
23425	23	122	23	2.6
23425	40	320		
...				

注: 第1列是病人编号，第2列是测试CD4的时刻（周），第3列是测得的CD4（乘以0.2个/ml），第4列是测试HIV的时刻（周），第5列是测得的HIV（单位不详）.

附件2: 193A数据（1300多名病人按照4种疗法服药大约每隔8周测试的CD4浓度）:

ID	疗法	年龄	时间	Log(CD4 count+1)
1	2	36.4271	0	3.1355
1	2	36.4271	7.5714	3.0445
1	2	36.4271	15.5714	2.7726
1	2	36.4271	23.5714	2.8332
1	2	36.4271	32.5714	3.2189
1	2	36.4271	40	3.0445
...				
1313	1	15.8412	0	4.9836
1313	1	15.8412	7.2857	4.1589
1313	1	15.8412	20	4.4067
1313	1	15.8412	27	3.5553
1313	1	15.8412	35	3.4657

注: 第1列是病人编号, 第2列是4种疗法的代码:

1 = 600mg zidovudine 与400mg didanosine按月轮换使用;

2 = 600mg zidovudine 加2.25mg zalcitabine;

3 = 600mg zidovudine 加400mg didanosine;

4 = 600mg zidovudine 加400mg didanosine 加400mg nevirapine.

第3列是病人年龄, 第4列是测试CD4的时刻（周）, 第5列是测得的CD4, 取值log(CD4+1).

需完成的问题:

(1) 利用附件1的数据, 预测继续治疗的效果, 或者确定最佳治疗终止时间（继续治疗指在测试终止后继续服药, 如果认为继续服药效果不好, 则可选择提前终止治疗）.

(2) 利用附件2的数据, 评价4种疗法的优劣（仅以CD4为标准）, 并对较优的疗法预测继续治疗的效果, 或者确定最佳治疗终止时间.

(3) 艾滋病药品的主要供给商对不发达国家提供的药品价格如下: 600mg zidovudine 1.60美元, 400mg didanosine 0.85美元, 2.25 mg zalcitabine 1.85美元, 400 mg nevirapine 1.20美元. 如果病人需要考虑4种疗法的费用, 对(2)中的评价和预测（或者提前终止）有什么改变.

- 数据类型: 纵向数据.

- 数据类型: 纵向数据.
- 预处理方案：

- 数据类型: 纵向数据.
- 预处理方案:
  - 删除/填补/缺失数据模型;

- 数据类型: 纵向数据.
- 预处理方案:
  - 删除/填补/缺失数据模型;
  - 分组: 年龄/病情;

- 数据类型: 纵向数据.
- 预处理方案:
  - 删除/填补/缺失数据模型;
  - 分组: 年龄/病情;
  - 标准化等数据变换;

- 数据类型: 纵向数据.
- 预处理方案:
  - 删除/填补/缺失数据模型;
  - 分组: 年龄/病情;
  - 标准化等数据变换;
  - 描述性分析.

问题一的求解

- 简单的线性回归方程：

$$\begin{cases} C_i(t) &= C_{i0} + \lambda_1 t \\ H_i(t) &= H_{i0} - \lambda_2 t \end{cases} \tag{11}$$

其中，C_{i0}表示病人i的CD4浓度初值，H_{i0}表示病人i的HIV浓度初值.

问题一的求解

- 简单的线性回归方程：

$$\begin{cases} C_i(t) &= C_{i0} + \lambda_1 t \\ H_i(t) &= H_{i0} - \lambda_2 t \end{cases} \tag{11}$$

其中, C_{i0}表示病人i的CD4浓度初值, H_{i0}表示病人i的HIV浓度初值.

- 抛物线回归模型:
模型(11)中λ_1表示药物疗效, 即CD4浓度的变化率. 考虑到生物个体具有抗药性, 用药时间延长, 药效会降低.
设λ_1是关于t减少的函数，如简单构造成$\lambda_1(t) = v_1 - u_1 t, (u_1 > 0, v_1 > 0)$. 同理构造$\lambda_2(t) = v_2 - u_2 t, (u_2 > 0, v_2 > 0)$, 模型(11)化为:

$$\begin{cases} C_i(t) - C_{i0} &= v_1 t - u_1 t^2 \\ H_i(t) - H_{i0} &= -v_2 t + u_2 t^2 \end{cases} \tag{12}$$

- 简单的指数回归模型：

假设CD4和HIV是分别孤立的, 即除了自身的分裂繁殖, 没有其他外界条件使二者浓度增减. 由此可得CD4和HIV浓度满足以下微分方程:

$$\begin{cases} \frac{dC(t)}{dt} &= \lambda_1 C(t) \\ \frac{dH(t)}{dt} &= -\lambda_2 H(t) \end{cases} \tag{13}$$

其中, λ_1, λ_2是对所有病人统一的待定常数.

- 改进的指数回归模型:

考虑到生物个体具有抗药性, 用药时间延长, 药效会降低, 其自然增长率不应为常数, 而应是关于时间t减少的函数, 于是可设:

$$\lambda_1(t)=\frac{\lambda_c}{(1+t)^2},\lambda_2(t)=\frac{\lambda_h}{(1+t)^2},$$

其中, λ_c,λ_h为常数.

将$\lambda_1(t),\lambda_2(t)$代入模型(13), 得改进的指数模型:

$$\begin{cases}\frac{dC(t)}{dt} &= \frac{\lambda_c}{(1+t)^2}C(t)\\ \frac{dH(t)}{dt} &= -\frac{\lambda_h}{(1+t)^2}H(t)\end{cases} \tag{14}$$

- CD4与HIV相互作用模型：

CD4不直接对HIV作用, 而是通过提高人体的免疫机能来抑制HIV病毒: HIV能直接使CD4感染而裂解, 使CD4数量急剧减少, HIV迅速增加, 导致AIDS发作.

类似于种群竞争模型, 利用微分方程建立模型:

$$\begin{cases} \frac{dC(t)}{dt} &= k_1 C(t) - k_2 H(t) \\ \frac{dH(t)}{dt} &= -k_3 H(t) \end{cases} \tag{15}$$

问题2的求解

基于抛物线模型的疗效分析

首先由附表2分别建立4种疗法的抛物线回归模型, 结合其一阶导数记入下表:

疗法(j)	回归函数$(C_j(t)-C_{j,i0})$	一阶导数	极值点
1	$a_1t^2-b_1t$	$2a_1t-b_1$	$T_1=b_1/(2a_1)$
2	$a_2t^2-b_2t$	$2a_2t-b_2$	$T_2=b_2/(2a_2)$
3	$a_3t^2-b_3t$	$2a_3t-b_3$	$T_3=b_3/(2a_3)$
4	$a_4t^2-b_4t$	$2a_4t-b_4$	$T_4=b_4/(2a_4)$

- 4种疗法的疗效比较:
  通过比较极值点时刻, 极值, 导数等综合评价.

- 4种疗法的疗效比较:
  通过比较极值点时刻, 极值, 导数等综合评价.
- 治疗效果的预测:
  (1) 在极值点处停药;
  (2) 给定阈值η, 在$\frac{dC(t)}{dt} = \eta$点处停药;
  (3) .....

基于核密度估计函数的疗效分析

考虑如下非参数纵向数据模型:

$$C_{ij} - C_{i0} = f(t_{ij}) + \epsilon_{ij}, i = 1, \cdots, n, j = 1, \cdots, m_i. \qquad (16)$$

其中随机误差ϵ_{ij}相互独立，均值为0.

采用核密度估计思想，引入核函数，确定出每个测量数据对于时刻t的响应变量预测值$\widehat{C}(t)$的贡献权重，建立出估计函数解析式，具体如下：

$$\widehat{C}(t) - C_{i0} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} w(t, t_{ij}, h)(C_{ij} - C_{i0}) / \sum_{i=1}^{n}\sum_{j=1}^{m_i} w(t, t_{ij}, h). \qquad (17)$$

其中，$w(t, t_{ij}, h) = K\{(t - t_{ij})/h\}$，$h$为核的带宽. h越大，估计函数越光滑.

后续分析同前.

其它可参考模型:

- 光滑样条非参数估计:
  极小化 $J(\lambda)=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{y_{ij}-s(t_{ij})\}^2+\lambda\int\{s''(t)\}^2dt.$

其它可参考模型:

- 光滑样条非参数估计:
  极小化 $J(\lambda)=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{y_{ij}-s(t_{ij})\}^2+\lambda\int\{s''(t)\}^2dt.$

- 半参数模型:

$$y_{ij}=f(x_{ij},\boldsymbol{\beta})+g(t_{ij})+\varepsilon_i(t_{ij}).$$

其它可参考模型:

- 光滑样条非参数估计:
  极小化 $J(\lambda)=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{y_{ij}-s(t_{ij})\}^2+\lambda\int\{s''(t)\}^2dt.$

- 半参数模型:

$$y_{ij}=f(x_{ij},\boldsymbol{\beta})+g(t_{ij})+\varepsilon_i(t_{ij}).$$

- 函数型数据模型:

$$y_i(t)=\mu_i(t)+\varepsilon_i(t).$$

其它可参考模型:

- 光滑样条非参数估计:
  极小化 $J(\lambda)=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{y_{ij}-s(t_{ij})\}^2+\lambda\int\{s''(t)\}^2dt.$

- 半参数模型:

$$y_{ij}=f(x_{ij},\boldsymbol{\beta})+g(t_{ij})+\varepsilon_i(t_{ij}).$$

- 函数型数据模型:

$$y_i(t)=\mu_i(t)+\varepsilon_i(t).$$

- ……

问题3的求解

- 考虑费用后疗效比较：
  引入单位费用疗效：$M_j = \frac{dC_j(t)}{dt}/n_j$，其中$\frac{dC_j(t)}{dt}$反映疗效，$n_j(j = 1, 2, 3, 4)$ 反映平均日花费. M_j越大说明综合效果越好.

问题3的求解

- 考虑费用后疗效比较：
  引入单位费用疗效：$M_j = \frac{dC_j(t)}{dt}/n_j$，其中$\frac{dC_j(t)}{dt}$反映疗效，$n_j(j = 1, 2, 3, 4)$ 反映平均日花费. M_j越大说明综合效果越好.
- 治疗效果预测：同上节讨论.

完成赛题应注意以下几点:

- 数据的合理使用;

完成赛题应注意以下几点:

- 数据的合理使用;
- 模型假设→模型建立、求解→模型检验→ $\cdots$ →模型应用;

完成赛题应注意以下几点:

- 数据的合理使用;
- 模型假设→模型建立、求解→模型检验→ $\cdots$ →模型应用;
- 模型多样性及相互比较;

完成赛题应注意以下几点:

- 数据的合理使用;
- 模型假设→模型建立、求解→模型检验→ $\cdots$ →模型应用;
- 模型多样性及相互比较;
- 模型和方法的连贯性.