第1章　模型概论

吴密霞

北京工业大学统计与数据科学系

E-mail:　*wumixia@bjut.edu.cn*

教材

- 吴密霞, 王松桂. 2024. 线性模型引论（第2版）, 科学出版社.

线性模型

线性模型是一类统计模型的总称, 它包括

- 线性回归模型
- 方差分析模型
- 协方差分析模型
- 线性混合效应模型...

在生物、医学、经济、管理、地质、气象、农业、工业、工程技术等许多领域的现象都可以用线性模型来近似描述.

几类常见的模型

- 线性回归模型
- 方差分析模型
- 协方差分析模型
- 混合效应模型
- 离散响应变量的广义线性模型
  - probit回归模型
  - logistic 模型

线性回归模型

在现实世界中, 存在着大量有一些依赖关系的变量, 其中一个或一些变量可以部分地且不很确切决定另一个变量的值, 例如

- 体重Y与身高X：当X大时, Y也倾向于大, 但由X 不能严格地决定Y;
- 城市生活用电量Y与气温X有很大的关系, 在夏天气温很高或冬天气温很低时, 用电量就高. 相反, 在春秋季节气温不高也不低, 用电量就相对少. 但不能由气温X准确地决定用电量Y.

变量之间的这种关系称为“相关关系”, 回归模型就是研究相关关系的一个有力工具.

线性回归模型

通常称Y为因变量或响应变量, X为自变量或解释变量. 设想Y的值由两部分组成:

- 由X能够决定的部分, 是X的函数, 记为$f(X)$;
- 由其他众多未加考虑的因素(包括随机因素)所产生的影响, 被看作随机误差, 记为e, 于是, 得

$$Y = f(X) + e, \tag{1.1}$$

这里, e为随机误差, $E(e) = 0$.

称模型(1.1)为回归模型. (关于“回归”一词后面作解释)

线性回归模型

若$f(X)$是线性的或者是近似线性的,

$$f(X) = \beta_0 + \beta_1 X,$$

则回归模型为

$$Y = \beta_0 + \beta_1 X + e. \tag{1.2}$$

称(1.2) 为线性回归模型, 称$\beta_0 + \beta_1 X$ 为线性回归方程, 其中

- β_0是直线的截距
- β_1 是直线的斜率, 也称为回归系数

β_0和β_1皆未知, 需通过观测数据来估计.

线性回归模型

假设有n组(X, Y)的观测值

$$(x_i, y_i), \qquad i = 1, \cdots, n.$$

如果Y与X 有回归关系(1.2), 则(x_i, y_i) 满足

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \cdots, n, \tag{1.3}$$

这里, e_i 为对应的随机误差. 应用适当统计方法可得到β_0和β_1的估计值$\hat{\beta}_0, \hat{\beta}_1$. 于是经验回归方程

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X.$$

“经验”表示这个回归直线是基于n次观测数据(x_i, y_i)估计所得, 非真实的

例1.1.1

肥胖是现代社会人们普遍关注的一个重要问题. 假设X表示身高(cm), Y表示体重(kg). 假设Y与X之间具有线性回归关系(1.2), 这里误差e表示除了身高X之外, 所有影响体重Y的其他因素, 例如遗传因素、饮食习惯、体育锻炼等.

某研究所由随机抽取的n人的身高和体重数据$(x_i, y_i), i = 1, \cdots, n$, 估计得$\hat{\beta}_0 = 40, \hat{\beta}_1 = 0.6$. 于是得经验回归方程

$$Y = -40 + 0.6X.$$

该方程在一定程度上描述了体重与身高的相关关系. 如对于一个身高160cm的人, 可预测他的体重大致为 56kg.

例1.1.2

我们知道, 一个公司的商品销售量Y与其广告费X_1有密切关系, 一般来说在其他因素(如产品质量等)保持不变的情况下, 用在广告上的费用越高, 它的商品销售量也就会愈多. 但这也只是一种相关关系.

某公司根据过去一段时间的销售记录得到经验回归方程

$$Y = 1608.5 + 20.1X.$$

经验回归方程表明：广告费X_1每增加一个单位, 该公司销售收入就增加20.1个单位.

如果该地区人口增加很快, 则人口总数X_2很可能也是影响销售量的一个重要因素. 根据记录的历史数据, 得到经验回归方程

$$Y = 320.3 + 18.4X_1 + 0.2X_2.$$

其表明:

- 当广告费X_1增加或人口总数X_2增加时, 商品销售量都增加;
- 当人口总数保持不变时, 广告费每增加1个单位, 销售量增加18.4个单位;
- 当广告费保持不变, 人口总数每增加1个单位, 销售量增加0.2个单位.

线性回归模型

在实际问题中, 影响因变量的主要因素往往很多. 假设因变量Y和$p-1$个自变量$X_1, \cdots, X_{p-1}$ 之间有如下关系:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + e, \tag{1.4}$$

其中β_0为常数项, $\beta_1, \cdots, \beta_{p-1}$为回归系数, e为随机误差.

对$(Y, X_1, \cdots, X_{p-1})$ 进行了n次观测, 得到n组观测值

$$(y_i, x_{i1}, \cdots, x_{i,p-1}), \quad i = 1, \cdots, n.$$

它们满足关系式

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1} + e_i, \quad i = 1, \cdots, n, \tag{1.5}$$

这里β_0为常数项, $\beta_1, \cdots, \beta_{p-1}$为回归系数, 皆未知, e_i为对应的随机误差, $E(e_i) = 0$. 记

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \ e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix},$$

线性回归模型

$$\mathbf{X}=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix},$$

则模型(1.5)的矩阵形式为

$$\boldsymbol{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{e}, \tag{1.6}$$

其中$\boldsymbol{y}$为$n\times 1$的观测向量, $\mathbf{X}$为$n\times p$已知设计矩阵, $\boldsymbol{\beta}$为未知参数向量, $\boldsymbol{e}$ 为$n\times 1$ 随机误差向量, $E(e_i)=0$.

这里“设计" 两字并不蕴含任何真正设计的含义, 只是习惯用法.

通常假设随机误差向量$\boldsymbol{e}$满足Gauss-Markov假设：

(a) 误差项具有等方差, 即 $\mathrm{Var}(e_i) = \sigma^2, \qquad i = 1, \cdots, n;$

(b) 误差是彼此不相关的, 即 $\mathrm{Cov}(e_i, e_j) = 0, \quad i \neq j.$

Gauss-Markov模型

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \qquad E(\boldsymbol{e}) = \mathbf{0}, \qquad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_n \tag{1.7}$$

这里$\mathrm{Cov}(\boldsymbol{e})$ 表示随机向量$\boldsymbol{e}$ 的协方差阵.

注 Gauss-Markov 假设要求e_i 等方差、不相关的条件, 这些要求有时显得严厉些.

线性回归模型

- 不同次观测序列相关.

  如在经济问题中, 一种最简单的自相关关系, 误差为一阶自回归:

$$e_i = \varphi e_{i-1} + \varepsilon_i, \qquad |\varphi| < 1,$$

  其中$\varepsilon_1, \cdots, \varepsilon_n$ 独立同分布, $E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$. 于是

$$\mathrm{Cov}(\boldsymbol{e}) = \frac{\sigma_\varepsilon^2}{1-\varphi^2} \begin{pmatrix} 1 & \varphi & \cdots & \varphi^{n-1} \\ \varphi & 1 & \cdots & \varphi^{n-2} \\ \vdots & \vdots & & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \cdots & 1 \end{pmatrix}. \tag{1.8}$$

线性回归模型

- 模型非线性的, 但经过适当变换, 可化为线性模型.

  如在经济学中著名的Cobb-Douglas生产函数：

$$Q_t = aL_t^b K_t^c,$$

  这里Q_t, L_t和K_t分别为t年的产值、劳力投入量和资金投入量. 在上式两边取自然对数, 得

$$\ln(Q_t) = \ln(a) + b\ln(L_t) + c\ln(K_t)\ .$$

  记$y_t = \ln(Q_t)$, $x_{t1} = \ln(L_t)$, $x_{t2} = \ln(K_t)$, $\beta_0 = \ln(a)$, $\beta_1 = b$, $\beta_2 = c$. 加上误差项, 便得到线性模型

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t.$$

线性回归模型

- 模型关于自变量非线性, 但关于未知参数线性.

  如多个自变量的多项式模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + e.$$

注 线性模型中“线性”实质上是指Y关于未知参数β_i的关系是线性的.

注 任何光滑函数都可用足够高阶的多项式来逼近, 故当因变量Y和诸自变量之间的关系不是线性关系时, 可用多元多项式来近似.

该类模型往往出现在化学工程领域的研究中, 其目的是求诸自变量的一个组合, 使得因变量Y 达到最大或最小. 这类问题称为响应曲面设计.

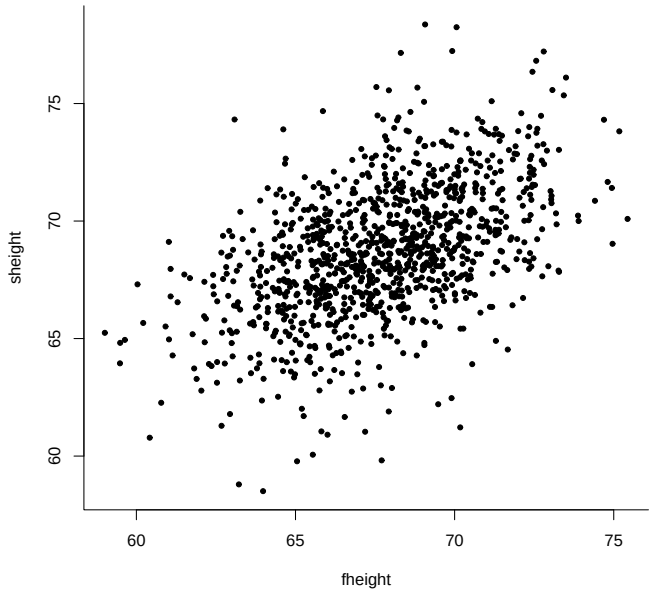
“回归"一词的由来

“回归"英文为“regression", 是由英国著名生物学家兼统计学家Galton在研究人类遗传问题时提出的. 为了研究父代与子代身高的关系, Galton收集了1078对父亲及其一子的身高数据.

- 用X表示父亲身高, Y表示儿子身高, 单位为英寸.
- R软件中可直接调用这个数据集，调用该数据集和画父子身高数据(x_i, y_i)散点图的程序如下：

```
install.packages("UsingR")
library(UsingR)
data(father.son)
plot(sheight   fheight, data=father.son, bty="l", pch=20)
```

“回归"一词的由来



- 散点图大致呈直线状. 总的趋势是父亲的身高X增加时, 儿子的身高Y也倾向于增加, 这与我们的常识是一致的.

“回归"一词的由来

- Galton深入分析, 发现了一个很有趣的现象—回归效应.
  - 这1078个父代平均身高、子代平均身高分别为为$\bar{x} = 68, \quad \bar{y} = 69$. 说明子代身高平均增加了1英寸.

自然推想

若父亲身高为x, 他儿子的平均身高y大致应为$x + 1$, 即$y \approx x + 1$.

但Galton的仔细研究所得结论与此大相径庭. 他发现

- 当父亲身高为72英寸时(高于平均), 他们儿子平均身高仅为71英寸. 不但达不到预期的72+1=73英寸, 反而比父亲身高低了1英寸.
- 若父亲身高为64英寸(矮于平均), 他们儿子平均身高为67英寸, 竟比预期的64+1=65英寸高出了2英寸.

“回归"一词的由来

这个现象不是个别的, 它反映了一个一般规律:

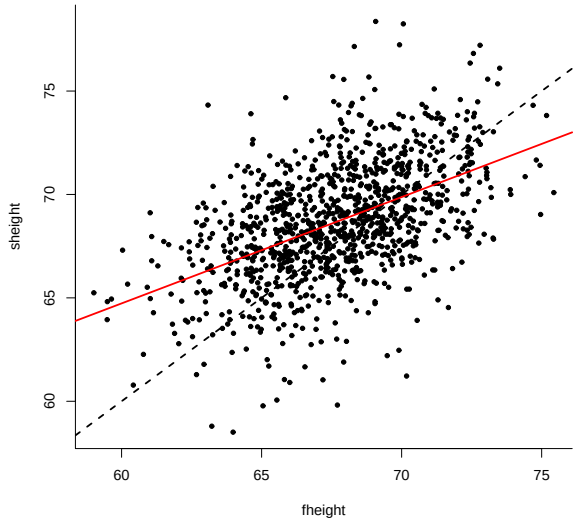
- 身高超过平均值$\bar{x} = 68$英寸的父亲, 他们儿子的平均身高将低于父亲的平均身高.
- 身高低于平均身高$\bar{x} = 68$英寸的父亲, 他们儿子的平均身高将高于父亲的平均身高.

Galton 对这个一般结论的解释

大自然具有一种约束力, 使人类身高的分布在一定时期内相对稳定而不产生两极分化, 这就是所谓的回归效应.

通过这个例子, Galton引进了“回归”一词. 用他的数据, 可以计算出儿子身高Y与父亲身高X的经验关系：$Y = 34 + 0.5X$.

“回归"一词的由来



- 它代表一条直线(图中红线), 称这条直线称为回归直线.
- 这个经验回归直线只反映了父子身高这两个变量相关关系中具有回归效应的一种特殊情况, 对更多的相关关系, 并非都是如此. 特别是涉及多个自变量的情况中, 回归效应便不复存在, 不过仍习惯性地沿用“回归" 这个词.

方差分析模型

- 线性回归模型
  - 所涉及的自变量一般来说都可以是连续变量,
  - 基本目的: 寻求因变量与自变量之间客观存在的依赖关系.
- 方差分析模型
  - 它的自变量是示性变量, 这种变量往往表示某种效应的存在与否, 因而只能取0, 1两个值.
  - 基本目的: 比较两个或多个因素效应大小.

比较因素效应的统计分析在统计学上叫做方差分析, 对应地将这种模型称为方差分析模型.

几类常见的模型

- 线性回归模型
- 方差分析模型
- 协方差分析模型
- 混合效应模型
- 离散响应变量的广义线性模型
  - probit回归模型
  - logistic 模型

例1.2.1 单向分类(one-way classification)模型

$$y_{ij} = \mu + \alpha_i + e_{ij}, \qquad i = 1, \cdots, a, \quad j = 1, \cdots, n_i, \tag{1.9}$$

这里μ称为总平均, α_i表示第i种处理的效应, e_{ij}表示随机误差, 其均值为0, 方差都相等, 彼此互不相关.

如比较三种药治疗某种疾病的效果, 药效度量指标为Y. 假设采用双盲实验法, 对每种药各有n个人服用, 记y_{ij}为服用第i种药的第j个患者的药效测量值, 则感兴趣的只有药品一个因素, 它有三个不同的品种, 称这三个品种为因子的水平或“处理", 即上模型中$a = 3, n_i = n$.

方差分析模型

模型可写为

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n} \\ \vdots \\ y_{a1} \\ \vdots \\ y_{an} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1n} \\ \vdots \\ e_{a1} \\ \vdots \\ e_{an} \end{pmatrix}.$$

用$\boldsymbol{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$和$\boldsymbol{e}$分别表示上式中的四个向量或矩阵, 则

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{1.10}$$

注 $\mathrm{rk}(\mathbf{X}) = a - 1 < a$. 设计阵列降秩是方差分析模型的一个特点.

方差分析模型

假设在一次生产实验中, 影响产品质量指标Y的有两个因素A和B. 设因素A有a个水平, 因素B 有b个水平. 记y_{ij}表示在因素A的第i个水平, 因素B的第j个水平时生产的产品质量测量值. 则y_{ij}可分解为

例1.2.2 两向分类(two-way classification)模型

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \qquad i = 1, \cdots, a, \quad j = 1, \cdots, n_i, \tag{1.11}$$

这里μ仍为总平均, α_i为因素A的第i个水平的效应, β_j为因素B的第j个水平的效应, e_{ij}为随机误差.

特例：随机区组设计模型

例 一农业实验中心从外地引进三种优良麦种, 在大面积种植之前, 先进行小范围试验以便选出适合本地气候条件的麦种.

- 把这三种小麦种植的施肥、浇水等条件控制在相同的状态；
- 对很难保证用于实验的土地肥沃程度, 随机区组设计：
  - 把实验用的土地分成若干小块(称作区组, block), 使每一小块土地肥沃程度基本上一样.
  - 再把每一区组分成3个更小的块(称作试验单元).
  - 将每种小麦完全是随机地种在各区组中一个试验单元. 试验单元的小麦产量 $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$, 其中$\alpha_i$为第$i$种小麦(即处理, treatment)的效应, β_j是第j个区组的效应.

方差分析模型

在试验设计中, 区组是一个很重要的概念. 两向分类模型往往也被称为随机区组设计模型, 并把α_i和β_j 分别泛称为 处理效应、区组效应.

通常主要感兴趣的是比较处理效应, 引入区组是为了缩小分析误差.

例 假设用a 种工艺加工一些产品, 现在要比较这a种工艺的优劣.

- 用y_{ij}表示第i种工艺加工的第j 件产品质量, α_i为第i 种工艺的效应,可用单向分类模型;
- 但若用b 台设备去检测它们的质量, 那么就应把这b 台设备的差异考虑进去. 这样b 台设备就成了区组, β_j 是第j 台设备的效应.

几类常见的模型

- 线性回归模型
- 方差分析模型
- 协方差分析模型
- 混合效应模型
- 离散响应变量的广义线性模型
  - probit回归模型
  - logistic 模型

协方差分析模型

- 线性回归模型
  - 自变量是连续变量, 设计阵$\mathbf{X}$的元素x_{ij} 可取连续值.
- 方差分析模型
  - 自变量是属性因子, 设计阵$\mathbf{X}$的元素x_{ij}只能取0和1 两个值.
- 协方差分析模型
  - 上述两种模型的混合. 模型中的自变量既有属性因子又有数量因子. 设计矩阵由两部分组成, 一部分以0和1两个数为元素, 而另一部分的元素可取连续值.
  - 可以看作由方差分析模型和线性回归模型的设计矩阵组拼而成.

协方差分析模型

经典例子

假定试验者用几种饲料喂养小猪, 并以小猪的生长速度(即小猪体重增加量)来比较饲料的催肥效果, 这是一个单向分类问题. 要求除饲料外, 其余因素应该尽量控制在相同条件之下. 但这里参与试验的小猪初始体重不同, 可能对生长速度有一定影响.

为了消除这种影响, 可以采取两种方法:

- 选择体重都一样的小猪来做试验. (在实际中困难很大)
- 设法把小猪初始体重的影响消除. 办法: 将其引入方差分析模型, 采用协方差分析.

协方差分析模型

例1.3.1

试验者欲比较两种饲料的催肥效果, 用每种饲料喂养三头猪.记y_{ij}为喂第i种饲料的第j头猪的体重增加量, 则y_{ij}可分解为

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + e_{ij}, \qquad i = 1, 2, \quad j = 1, 2, 3, \tag{1.12}$$

这里μ为总平均, α_i为第i种饲料的效应, x_{ij}为喂第i种饲料的第j头猪的初始体重, γ为协变量的系数, 即回归系数.

猪的饲料分几个品种, 是属性因子, 称为方差分量. 由于小猪的初始体重是试验者难以很好地控制而进入试验的, 故称其为协变量(或伴随变量), 它是连续变量.

协方差分析模型

若记 $\boldsymbol{y} = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23})'$，

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & x_{11} \\ 1 & 1 & 0 & x_{12} \\ 1 & 1 & 0 & x_{13} \\ 1 & 0 & 1 & x_{21} \\ 1 & 0 & 1 & x_{22} \\ 1 & 0 & 1 & x_{23} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \gamma \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{pmatrix},$$

则模型(1.12)具有形式

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

特点: 设计阵$\mathbf{X}$部分列元素只取0或1, 剩余列元素则取连续值

几类常见的模型

- 线性回归模型
- 方差分析模型
- 协方差分析模型
- 混合效应模型
- 离散响应变量的广义线性模型
  - probit回归模型
  - logistic 模型

线性混合效应模型

例1.4.1

以研究人的血压在一天内的变化规律为例. 在一天内选择a个时间点测量被观测者的血压, 假设观测了b个人, 用y_{ij}表示第i个时间点的第j个人的血压, α_i为第i个时间点的效应, β_j为第j个人的个体效应, 则

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, b,$$

这里α_i是固定效应, 而β_j是否随机取决于研究兴趣.

- 固定效应 若这b个人是感兴趣的特定的b个人
- 随机效应 若研究的兴趣只是比较不同时间点人的血压高低上, 被观测的b个人是随机抽取的.

例1.4.1 面板数据(Panel data) 模型

在计量经济学中,假设我们对N个个体(如个人, 家庭, 公司, 城市, 国家或区域等) 进行了T个时刻的观测,

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \xi_i + \varepsilon_{it}, \qquad i = 1, \cdots, N, \quad t = 1, \cdots, T, \tag{1.13}$$

其中y_{it}表示第i个个体第t个时刻的某项经济指标, $\boldsymbol{x}_{it}$是$p \times 1$已知向量, 刻画了第i个个体在时刻t的一些自身特征, ξ_i是第i个个体的个体效应, ε_{it} 是随机误差.

若目的是研究整个市场的运行规律, 这N个个体是从总体中抽取的随机样本,而非特定关心的个体, 这时个体效应是随机的.

线性混合效应模型

针对模型(1.13), 记 $\boldsymbol{y} = (y_{11}, \cdots, y_{1T}, \cdots, y_{N1}, \cdots, y_{NT})'$,

$$\mathbf{X} = (x_{11}, \cdots, x_{1T}, \cdots, x_{N1}, \cdots, x_{NT})', \quad \mathbf{U}_1 = \mathbf{I}_N \otimes \mathbf{1}_T,$$

$$\boldsymbol{\xi} = (\xi_1, \cdots, \xi_N)', \quad \boldsymbol{\varepsilon} = (\varepsilon_{11}, \cdots, \varepsilon_{1T}, \cdots, \varepsilon_{N1}, \cdots, \varepsilon_{NT})'.$$

假设$\mathrm{Var}(\xi_i) = \sigma_\xi^2$, $\mathrm{Var}(\varepsilon_{it}) = \sigma_\varepsilon^2$, 所有$\xi_i$和$\varepsilon_{it}$都不相关, 则

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}_1\boldsymbol{\xi} + \boldsymbol{\varepsilon},$$

$$\mathrm{Cov}(\boldsymbol{y}) = \sigma_\xi^2 \mathbf{U}_1\mathbf{U}_1' + \sigma_\varepsilon^2 \mathbf{I}_{NT} = \sigma_\xi^2 (\mathbf{I}_N \otimes \mathbf{J}_T) + \sigma_\varepsilon^2 \mathbf{I}_{NT},$$

这里, $\otimes$ 为Kronecker乘积(见第2章), $\mathbf{1}_n$ 表示$n \times 1$向量,它的所有元素均为1, $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n'$.

线性混合效应模型

混合效应模型一般形式

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}_1\boldsymbol{\xi}_1 + \mathbf{U}_2\boldsymbol{\xi}_2 + \cdots + \mathbf{U}_k\boldsymbol{\xi}_k, \tag{1.14}$$

其中, $\mathbf{X}$和$\mathbf{U}_i$为已知设计阵, $\boldsymbol{\beta}$为固定效应, $\boldsymbol{\xi}_i$为随机效应. 假设

$$E(\boldsymbol{\xi}_i) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{\xi}_i) = \sigma_i^2\mathbf{I}_{q_i}, \quad \mathrm{Cov}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = \mathbf{0}, \quad i \neq j,$$

于是

$$E(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta}, \qquad \mathrm{Cov}(\boldsymbol{y}) = \sum_{i=1}^{k} \sigma_i^2\mathbf{U}_i\mathbf{U}_i', \tag{1.15}$$

称σ_i^2为方差分量, 称(1.14)为方差分量模型.

几类常见的模型

- 线性回归模型
- 方差分析模型
- 协方差分析模型
- 混合效应模型
- 离散响应变量的广义线性模型
  - probit回归模型
  - logistic 模型

离散响应变量模型

响应变量Y可能为分类变量或计数变量. 如

- 考虑是否结婚、是否生二胎，响应变量是二分类(0-1) 变量;
- 考虑出行交通工具的选择(1. 步行, 2. 电动车, 3. 汽车, 4.地铁)时, 响应变量是无序多分类变量;
- 考虑顾客对酒店环境的评价(1. 不满意, 2. 一般, 3.满意, 4. 非常满意)时, 响应变量是有序多分类变量;
- 在检验某种治疗癫痫药物的药效时, 记录两个月内治疗组和对照组病人癫痫发作的次数就是一个计数变量等.

离散响应变量模型

妊娠持续时间模型(Kutner, 2004)为例来显示二分类响应变量的一种建模思想.

随机抽取n名孕妇，记录她们怀孕时间和孕期酗酒程度

$$(y_i^c, x_i), \quad i = 1, \cdots, n.$$

假设母亲酗酒程度(X)对其怀孕时间(Y^c)的影响满足模型:

$$y_i^c = \beta_0^c + \beta_1^c x_i + e_i^c.$$

若感兴趣的是母亲酗酒是否会导致婴儿早产, 此时响应变量

$$y_i = \begin{cases} 1, & y_i^c \leq 38 \text{ 周}, \quad (\text{早产}), \\ 0, & y_i^c > 38 \text{ 周}, \quad (\text{足月}). \end{cases}$$

离散响应变量模型

记F_{e^c} 为模型随机误差e_i^c的分布函数. 于是响应变量y_i的均值为

$$\pi(\boldsymbol{x}_i) = E(y_i) = P(e_i^c \leq 38 - \beta_0^c - x_i\beta_1^c) = F_{e^c}(38 - \beta_0^c - x_i\beta_1^c).$$

- 若模型误差$e_i^c \sim N(0, \sigma^2)$, 则

$$\pi(\boldsymbol{x}_i) = \Phi\left(\frac{38 - \beta_0^c}{\sigma} - \frac{\beta_1^c}{\sigma}x_i\right),$$

这里, Φ为标准正态分布函数. 记$\beta_0 = \frac{d-\beta_0^c}{\sigma}$, $\beta_1 = -\frac{\beta_1^c}{\sigma}$.
得probit回归模型

$$\text{probit}(\pi(x_i)) = \Phi^{-1}(\pi(x_i)) = \beta_0 + x_i\beta. \tag{1.14}$$

- 若e_i^c服从logistic 分布, 则

  e_i^c可表示为$e_i^c = \sigma\pi\varepsilon_i/\sqrt{3}$, 其中, ε_i服从标准的logistic分布, 分布函数为

$$F_\varepsilon(\varepsilon_i) = \frac{\exp(\varepsilon_i)}{1+\exp(\varepsilon_i)} = \frac{1}{1+\exp(-\varepsilon_i)} = h(\varepsilon_i). \tag{1.15}$$

  记$\beta_0 = \pi(d-\beta_0^c)/(\sqrt{3}\sigma)$, $\beta_1 = -\pi\beta_1^c/(\sqrt{3}\sigma)$. 于是,

$$\pi(x_i) = P(y_i = 1) = P(\varepsilon_i \le \beta_0 + \boldsymbol{x}_i\beta_1) = \frac{1}{1+\exp\{-\beta_0 - x_i\beta_1\}}.$$

  得logistic回归模型:

$$\text{logit}(\pi(x_i)) = \beta_0 + x_i\beta_1, \tag{1.16}$$

  其中, $\text{logit}(\pi(x_i)) = \ln\left(\pi(x_i)/(1-\pi(x_i))\right)$.

离散响应变量模型

注 差e_i^c分布的不同假设, 就会得到不同的二分类模型.

另外, 为了研究这多个危险因子对妊娠持续时间的影响，研究者又将妊娠持续时间小于38周细分为两类：

- 不足36周（早产）
- 36周到37周之间（介于早产和足月之间）

此时因变量就变成了三分类变量，而且这三类之间有顺序.

关于此类数据的建模将在第10章给出

小结

- 初步了解线性模型中常见的几类模型特点:
  - 线性回归模型
  - 方差分析模型
  - 协方差分析模型
  - 混合效应模型
- 初步了解二分类散响应变量的建模思想