数学建模中的统计方法

Statistical Methods in Mathematical modeling

马学俊

献给我的家人、恩师和所有在学术道路上帮助我的人

马学俊, 副教授，苏州大学数学科学学院统计系，主要从事海量数据分析、高维数据分析、统计计算、非参数回归等统计模型及其应用等研究。个人主页 https://xuejunma.github.io.

目录

第一部分

统计方法

第 1 章 多元线性回归

1.1 Credit 数据

Caredit 数据包含 10 自变量个变量:

- Income: 收入 (单位是千美元)
- Limit: 信用额度
- Rating: 信用级别 (连续变量)
- Cards : 信用卡数量
- Age: 年龄
- Education: 受教育年限
- Gender: 性别
- Student: 学生
- Married: 婚姻状况
- Ethnicity: 种族 (白种人、非裔美国人、亚洲人)

目的: 分析哪些因素影响个人平均信用卡债务 Balance（因变量）。

需要强调 `Ethnicity` 是分类变量，有三个类别。

Credit

```
> rm(list=ls())
> Credit_lm  <- read.csv("/Users/yinuo/Desktop/数学建模/Credit.csv")
> head(Credit_lm)
  X  Income Limit Rating Cards Age Education Gender Student Married Ethnicity Balance
1 1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian     333
2 2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian     903
3 3 104.593  7075    514     4  71        11   Male      No      No     Asian     580
4 4 148.924  9504    681     3  36        11 Female      No      No     Asian     964
5 5  55.882  4897    357     2  68        16   Male      No     Yes Caucasian     331
6 6  80.180  8047    569     4  77        10   Male      No      No Caucasian    1151
```

1.2 方法

1.2.1 多元线性回归模型的一般形式

设随机变量 y 与一般变量 $x_1, x_2, \ldots, x_p$ 的线性回归模型为:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

对于随机误差项假定:

$$\begin{cases} E(\epsilon) = 0 \\ var(\epsilon) = \sigma^2 \end{cases}$$

对于 n 组观测数据 $(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i), i = 1, 2, \ldots, n$，线性回归模型表示为：

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_2 \\ \quad \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \epsilon_n \end{cases}$$

写成矩阵的形式为

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

其中

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \text{(设计矩阵)}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

1.2.2 多元线性回归模型的基本假定

(1) 解释变量 $x_1, x_2, \ldots, x_p$ 是确定性变量，不是随机变量，且要求 $rank(\boldsymbol{X}) = p + 1 < n$。表明设计矩阵 $\boldsymbol{X}$ 中的自变量列之间不相关，$\boldsymbol{X}$ 是一满秩矩阵。

(2) 随机误差项具有 0 均值和等方差, 即

$$\begin{cases} E(\epsilon_i) = 0, \quad i = 1, 2, \ldots, n \\ cov(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2, i = j \\ 0, i \neq j \end{cases} \quad i, j = 1, 2, \ldots, n \end{cases}$$

这个假定称为 Gauss-Markov 条件

(3) 正态分布的假定条件为:

$$\begin{cases} \epsilon_i \sim N(0, \sigma^2) \\ \epsilon_1, \epsilon_2, \ldots, \epsilon_n \quad \text{相互独立} \end{cases}$$

矩阵形式表示为

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$$

在正态假定下：

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

$$E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$$

$$var(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$$

1.2.3 多元线性回归方程的解释

- y 表示空调机的销售量，
- x_1 表示空调机的价格，
- x_2 表示消费者可用于支配的收入。

建立二元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

在 x_2 保持不变时，$\frac{\partial E(y)}{\partial x_1} = \beta_1$

- β_1 可解释为在消费者收入 x_2 保持不变时，空调机价格 x_1 每增加一个单位，空调机销售量 y 的平均增加幅度。

在 x_1 保持不变时，$\frac{\partial E(y)}{\partial x_2} = \beta_2$

1.2.4 回归参数的估计

1.2.4.1 最小二乘估计

最小二乘估计要寻找 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$，使得

$$Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

$$= \min_{\beta_0, \beta_1, \beta_2, \ldots, \beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

$$\begin{cases} \frac{\partial Q}{\partial \beta_0}|_{\beta_0=\hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) \\ \frac{\partial Q}{\partial \beta_1}|_{\beta_1=\hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) x_{i1} \\ \frac{\partial Q}{\partial \beta_2}|_{\beta_2=\hat{\beta}_2} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) x_{i2} \\ \vdots \\ \frac{\partial Q}{\partial \beta_p}|_{\beta_p=\hat{\beta}_p} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) x_{ip} \end{cases}$$

整理后得到矩阵形式表示的正规方程组

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$$

移项得

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

当 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 存在时，即得到回归参数的最小二乘估计为

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

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad \Rightarrow \quad |\boldsymbol{X}'\boldsymbol{X}| \neq 0 \quad \Rightarrow \quad rank(\boldsymbol{X}'\boldsymbol{X}) = p+1$$
$$\Rightarrow \quad rank(\boldsymbol{X}) \geq p+1 \quad \Rightarrow \quad \boldsymbol{X}_{n\times(p+1)} \quad \Rightarrow \quad n \geq p+1$$

称

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$

为观测值 y_i 得回归拟合值。

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

称

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

为帽子矩阵, 其主对角线元素记为 h_{ii}.

- $tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p+1$。

$$tr(\boldsymbol{H}) = tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = tr((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = tr(\boldsymbol{I}_{p+1}) = p+1$$

- $\boldsymbol{H}^2 = \boldsymbol{H}$

回归残差向量

$$\boldsymbol{e} = (e_1, e_2, \cdots, e_n)' = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

$$\begin{aligned} D(\boldsymbol{e}) &= cov(\boldsymbol{e}, \boldsymbol{e}) \\ &= cov((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}, (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}) \\ &= (\boldsymbol{I}-\boldsymbol{H})cov(\boldsymbol{y}, \boldsymbol{y})(\boldsymbol{I}-\boldsymbol{H})' \\ &= \sigma^2(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{I}_n(\boldsymbol{I}-\boldsymbol{H})' \\ &= \sigma^2(\boldsymbol{I}-\boldsymbol{H}) \end{aligned}$$

得到

$$D(e_i) = (1-h_{ii})\sigma^2$$

又因为

$$E(\sum_{i=1}^{n} e_i^2) = \sum_{i=1}^{n} D(e_i) = (n-p-1)\sigma^2$$

可得

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-p-1} SSE = \frac{1}{n-p-1}(\boldsymbol{e}'\boldsymbol{e}) \\ &= \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2 \end{aligned}$$

1.2.4.2 回归参数得最大似然估计

$\boldsymbol{y}$ 得概率分布为

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$$

可以得到似然函数为

$$L = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}))$$

两边同时取对数似然

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$$

等价于使 $(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$ 达到最小，这又与普通最小二乘相同。

1.2.5 参数估计量的性质

- 性质 1：$\hat{\boldsymbol{\beta}}$ 是随机向量 $\boldsymbol{y}$ 的一个线性变换。

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

- 性质 2：$\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的无偏估计。

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{X\beta}+\boldsymbol{\epsilon}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

- 性质 3：$D(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

$$\begin{aligned} D(\hat{\boldsymbol{\beta}}) &= cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}) \\ &= cov((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'cov(\boldsymbol{y},\boldsymbol{y})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

- 性质 4：Gauss-Markov 定理
  在假定 $E(\boldsymbol{y}) = \boldsymbol{X\beta}, D(\boldsymbol{y}) = \sigma^2\boldsymbol{I}_n$ 时，$\boldsymbol{\beta}$ 的任一线性函数 $\boldsymbol{c}'\boldsymbol{\beta}$ 的最小方差线性无偏估计（BLUE）为 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$，其中，$\boldsymbol{c}$ 是任一维常数向量，$\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的最小二乘估计。
  (1) 取常数向量 $\boldsymbol{c}$ 的第 $j(j=0,1,\ldots,p)$ 个分量为 1，其余分量为 0，这时 G-M 定理表明最小二乘估计 $\hat{\beta}_j$ 是 β_j 的最小方差线性无偏估计。
  (2) 可能存在 $y_1, y_2, \ldots, y_n$ 的非线性函数，作为 $\boldsymbol{c}'\boldsymbol{\beta}$ 的无偏估计，比最小二乘估计 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 的方差更小。
  (3) 可能存在 $\boldsymbol{c}'\boldsymbol{\beta}$ 的有偏估计量，在某种意义（例如均方误差最小）下比最小二乘估计 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 更好。
  (4) 在正态假定下，$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{c}'\boldsymbol{\beta}$ 的最小方差无偏估计。也就是说，既不可能存在 $y_1, y_2, \ldots, y_n$ 的非线性函数，也不可能存在 $y_1, y_2, \ldots, y_n$ 的其它线性函数，作为 $\boldsymbol{c}'\boldsymbol{\beta}$ 的无偏估计，比最小二乘估计 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 方差更小。

1.2.6 回归方程的显著性检验

1.2.6.1 F 检验

原假设

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

利用总离差平方和的分解式

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

$$SST=SSR+SSE$$

构造 F 检验统计量如下

$$F=\frac{SSR/p}{SSE/(n-p-1)}$$

当原假设成立时，F 服从自由度为 $(p,n-p-1)$ 的 F 分布。

方差来源	自由度	平方和	均方	F 值	p 值
回归	p	SSR	$\frac{SSR}{p}$	$\frac{SSR/p}{SSE/(n-p-1)}$	$P(F>F)$
残差	$n-p-1$	SSE	$\frac{SSE}{n-p-1}$		
总和	$n-1$	SST			

1.2.6.2 t 检验

原假设

$$H_0:\quad \beta_j=0,\quad j=1,2,\ldots,p$$

已知

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

记

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}=(c_{ij}),\quad i,j=0,1,2,\ldots,p$$

由此可以构造 t 统计量

$$t_j=\frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$$

其中

$$\hat{\sigma}=\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}e_i^2}=\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

1.3 几个关键的问题

1.3.1 定性变量

首先讨论定性变量只取两类可能值的情况，例如研究粮食产量问题，y 为粮食产量，x 为施肥量，另外再考虑气候问题，分为正常年份和干旱年份两种情况，对这个问题的数量化方法是引入一个 $0-1$ 型变量 D，令：

$$D_i=1\text{表示正常年份}$$

$$D_i=0\text{表示干旱年份}$$

粮食产量的回归模型为

$$y_i=\beta_0+\beta_1x_i+\beta_2D_i+\epsilon_i$$

干旱年份的粮食平均产量为

$$E(y_i|D_i=0)=\beta_0+\beta_1 x_1$$

正常年份的粮食平均产量为：

$$E(y_i|D_i=1)=(\beta_0+\beta_2)+\beta_1 x_1$$

假设条件：干旱年份和正常年份回归直线的斜率 β_1 相等

某些场合定性自变量可能取多类值，例如某商厦策划营销方案，需要考虑销售额的季节性影响，季节因素分为春、夏、秋、冬 4 种情况。为了用定性自变量反应春、夏、秋、冬四季，我们初步设想引入如下 4 个 0 – 1 自变量：

$$\begin{cases} x_1=1, \text{春季} \\ x_1=0, \text{其他} \end{cases} \quad \begin{cases} x_2=1, \text{夏季} \\ x_2=0, \text{其他} \end{cases}$$

$$\begin{cases} x_3=1, \text{秋季} \\ x_3=0, \text{其他} \end{cases} \quad \begin{cases} x_4=1, \text{冬季} \\ x_4=0, \text{其他} \end{cases}$$

可是这样做却产生了一个新的问题，即 $x_1+x_2+x_3+x_4=1$，构成完全多重共线性。解决这个问题的方法很简单，我们只需去掉一个 0 – 1 型变量，只保留 3 个 0 – 1 型自变量即可。例如去掉 x_4，只保留 x_1、x_2、x_3。

对一般情况，一个定性变量有 k 类可能的取值时，需要引入 $k-1$ 个 0 – 1 型自变量。当 $k=2$ 时，只需要引入一个 0 – 1 型自变量即可。

图**??**的代码

```
> rm(list=ls())
> library(nnet)
> Credit_lm  <- read.csv("/Users/yinuo/Desktop/数学建模/Credit.csv")
> head(Credit_lm)
> data <- Credit_lm[-1]
> attach(data)
> stu <- nnet::class.ind(data$Student)
> Student1 <- stu[,2]# 是学生为 1
> gen <- nnet::class.ind(data$Gender)
> Gender1 <- gen[,2]# 性别女为 1
> mar <- nnet::class.ind(data$Married)
> Married1 <- mar[,2]# 已婚为 1
> eth <- nnet::class.ind(data$Ethnicity)
> Ethnicity1 <- eth[,1]# 是否为非裔美国人
> Ethnicity2 <- eth[,2]# 是否为亚裔
> fit <- lm(Balance~Income+Limit+Rating+Cards+Age+Education+Gender1+Student1+Married1+Ethnicity1+E
> summary(fit)

Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
    Education + Gender1 + Student1 + Married1 + Ethnicity1 +
    Ethnicity2, data = data)

```

```
Residuals:
    Min      1Q  Median      3Q     Max
-161.64  -77.70  -13.49   53.98  318.20

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -469.10085   34.70899 -13.515  < 2e-16 ***
Income        -7.80310    0.23423 -33.314  < 2e-16 ***
Limit          0.19091    0.03278   5.824 1.21e-08 ***
Rating         1.13653    0.49089   2.315   0.0211 *
Cards         17.72448    4.34103   4.083 5.40e-05 ***
Age           -0.61391    0.29399  -2.088   0.0374 *
Education     -1.09886    1.59795  -0.688   0.4921
Gender1      -10.65325    9.91400  -1.075   0.2832
Student1     425.74736   16.72258  25.459  < 2e-16 ***
Married1      -8.53390   10.36287  -0.824   0.4107
Ethnicity1   -10.10703   12.20992  -0.828   0.4083
Ethnicity2     6.69715   12.12244   0.552   0.5810
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 98.79 on 388 degrees of freedom
Multiple R-squared:  0.9551,    Adjusted R-squared:  0.9538
F-statistic: 750.3 on 11 and 388 DF,  p-value: < 2.2e-16
```

1.3.2 异常值与强影响点

异常值分为三种情况：

- 一种是关于因变量 y 异常；Outlier
- 另一种是关于自变量 x 异常：高杠杆点 High-leverage point
- 关于模型异常：强影响点 Influence point

1.3.2.1 关于因变量 y 的异常值

在残差分析中，认为超过 $\pm 3\hat{\sigma}$ 的残差为异常值。标准化残差

$$ZRE_i = \frac{e_i}{\hat{\sigma}}$$

学生化残差

$$SRE_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

h_{ii} 为 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 的主对角线元素。

当观测数据中存在关于 y 的异常观测值时，普通残差、标准化残差、学生化残差这三种残差都不再适用。这是由于异常值把回归线拉向自身，使异常值本身的残差减小，而其余观测值的残差增大，这时回归标准差 $\hat{\sigma}$ 也会增大，因而用传统的“3σ”准则不能正确分辨出异常值。解决这个问题的方法是改用删除残差。

删除残差的构造思想是：在计算第 i 个观测值的残差时，用删除掉的第 i 个观测值的其余 $n-1$ 个观测值拟合回归方程，计算出第 i 个观测值的删除拟合值 $\hat{y}_{(i)}$，这个删除拟合值与第 i 个值无关，不受第 i 个值是否为异常值的影响，由此定义第 i 个观测值的删除残差为

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

可以证明

$$e_{(i)} = \frac{e_i}{1-h_{ii}}$$

进一步，可以给出第 i 个观测值的删除学生残差，记为 $SRE_{(i)}$。

$$SRE_{(i)} = SRE_i\left(\frac{n-p-2}{n-p-1-SRE_i^2}\right)^{\frac{1}{2}}$$

$|SRE_{(i)}| > 3$ 的观测值即判定为异常值。

1.3.2.2 自变量 x 的异常值对回归的影响 (高杠杆点 High-leverage point)

- 在 $D(e_i) = (1-h_{ii})\sigma^2$ 中，h_{ii} 为帽子矩阵中主对角线的第 i 个元素，它是调节 e_i 方差大小的杠杆，因而称 h_{ii} 为第 i 个观测值的杠杆值。类似于一元线性回归，多元线性回归的杠杆值 h_{ii} 也表示自变量的第 i 次观测与自变量平均值之间距离的远近。
- 较大的杠杆值的残差偏小，这是因为杠杆值大的观测点远离样本中心，能够把回归拉向自身，因而把杠杆值大的样本点称为强影响点。

- $tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p+1$，则杠杆值的平均值为

  $$\bar{h} = \frac{1}{n}h_{ii} = \frac{p+1}{n}$$

  一个杠杆值 h_{ii} 大于 2 倍或者 3 倍的 $\bar{h}$，就认为是大的。

1.3.2.3 强影响点 (Influence point)

- 虽然强影响点并不总是 y 的异常值点，不能单纯根据杠杆值 h_{ii} 的大小判断强影响点是否异常，但是我们对强影响点应该有足够的重视。
- 为此引入库克距离，用来判断强影响点是否为 y 的异常值点。库克距离的计算公式为：

  $$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

- 库克距离反映了杠杆值 h_{ii} 与残差 e_i 的综合效应。对于库克距离，判断其大小的方法比较复杂，一个粗略的标准是:
  - 当 $D_i < 0.5$ 时，认为不是异常值点，
  - 当 $D_i > 1$ 时，认为是异常值点。

1.3.3 多重共线性的情形及其处理

- 如果存在不全为 0 的 $p+1$ 个数 $c_0, c_1, c_2, \cdots, c_p$，使得

  $$c_0 + c_1x_{i1} + c_2x_{i2} + \cdots + c_px_{ip} = 0, i = 1, 2, \cdots, n \tag{6.1}$$

  则称自变量 $x_1, x_2, \cdots, x_p$ 之间存在着完全多重共线性。
- 在实际经济问题中完全的多重共线性并不多见，常见的是 (6.1) 式近似成立的情况，即存在不

全为 0 的 $p+1$ 个数 $c_0, c_1, c_2, \cdots, c_p$，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} \approx 0, i = 1, 2, \cdots, n \tag{6.1}$$

称自变量 $x_1, x_2, \cdots, x_p$ 之间存在着多重共线性 (Multi-collinearity), 也称为复共线性。

1.3.3.1 多重共线性产生的经济背景和原因

- 当我们所研究的经济问题涉及到时间序列资料时, 由于经济变量随时间往往存在共同的变化趋势, 使得它们之间就容易出现共线性。
  - 例如, 我们要研究我国居民消费状况, 影响居民消费的因素很多, 一般有职工平均工资、农民平均收入、银行利率、全国零售物价指数、国债利率、货币发行量、储蓄额、前期消费额等, 这些因素显然既对居民消费产生重要影响, 它们之间又有着很强的相关性。
- 许多利用截面数据建立回归方程的问题常常也存在自变量高度相关的情形。
  - 例如, 我们以企业的截面数据为样本估计生产函数, 由于投入要素资本 K, 劳动力投入 L, 科技投入 S, 能源供应 E 等都与企业的生产规模有关, 所以它们之间存在较强的相关性。

1.3.3.2 多重共线性对回归模型的影响

设回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

存在完全的多重共线性，即对设计矩阵 $\boldsymbol{X}$ 的列向量存在不全为 0 的一组数 $c_0, c_1, c_2, \cdots, c_p$，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} = 0, i = 1, 2, \cdots, n$$

- 设计矩阵 $\boldsymbol{X}$ 的秩 $rank(\boldsymbol{X}) < p+1$, 此时 $|\boldsymbol{X}'\boldsymbol{X}| = 0$, 正规方程组 $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ 的解不唯一，
- $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 不存在，回归参数色最小二乘估计表达式 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 不成立。

对非完全共线性, 存在不全为零的一组数 $c_0, c_1, c_2, \cdots, c_p$，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} \approx 0, i = 1, 2, \cdots, n$$

此时设计矩阵 $\boldsymbol{X}$ 的秩 $rank(\boldsymbol{X}) = p+1$ 虽然成立, 但是 $|\boldsymbol{X}'\boldsymbol{X}| \approx 0$,

- $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ 的估计精度很低
  - $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 的对角元素很大，$\hat{\boldsymbol{\beta}}$ 的方差矩阵 $D(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 的对角元素很大，而 $D(\hat{\boldsymbol{\beta}})$ 的对角元素即 $var(\hat{\beta}_0), var(\hat{\beta}_1), \cdots, var(\hat{\beta}_p)$
- 虽然用普通最小二乘估计能得到 $\boldsymbol{\beta}$ 的无偏估计，但估计量 $\hat{\boldsymbol{\beta}}$ 的方差很大
- 不能正确判断解释变量对被解释变量的影响程度
- 甚至会导致估计量的经济意义无法解释。

对于二元回归模型，做 y 对两个自变量 x_1, x_2 的线性回归, 假定 y 与 x_1, x_2 都已经中心化，此时回归常数项为零，回归方程为

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

记 $L_{11} = \sum_{i=1}^n x_{i1}^2, L_{12} = \sum_{i=1}^n x_{i1}x_{i2}, L_{22} = \sum_{i=1}^n x_{i2}^2$，则 x_1 与 x_2 相关系数为

$$r_{12} = \frac{L_{12}}{\sqrt{L_{11}L_{22}}}$$

$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ 的协方差矩阵为

$$cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} L_{12} & L_{12} \\ L_{12} & L_{22} \end{pmatrix}$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{|\boldsymbol{X}'\boldsymbol{X}|}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix} = \frac{1}{L_{11}L_{22} - L_{12}^2}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix}$$

$$= \frac{1}{L_{11}L_{22}(1 - r_{12}^2)}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix}$$

由此可得

$$var(\hat{\beta}_1) = \frac{\sigma^2}{(1 - r_{12}^2)L_{11}}$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{(1 - r_{12}^2)L_{22}}$$

可知，随着自变量 x_1 与 x_2 的相关性增强，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差将逐渐增大，当 x_1 与 x_2 完全相关时，$r = 1$，方差将变为无穷大。

- 当给不同的 r_{12} 值时, 由下表可看出方差增大的速度。
- 为了方便，我们假设 $\sigma^2/L_{11} = 1$, 相关系数从 0.5 变为 0.9 时, 回归系数的方差增加了 295%, 相关系数从 0.5 变为 0.95 时, 回归系数的方差增加了 671%。

1.3.3.3 多重共线性的诊断

方差扩大因子法 对自变量做中心标准化，则 $\boldsymbol{X}^{*'}\boldsymbol{X}^* = (r_{ij})$ 为自变量的相关阵，记

$$\boldsymbol{C} = (c_{ij}) = (\boldsymbol{X}^{*'}\boldsymbol{X}^*)^{-1}$$

称其主对角线元素 $VIF_j = c_{jj}$ 为自变量 x_j 的方差扩大因子(Variance Inflation Factor, 简记为 VIF)。根据书（3.31）式可知，

$$var(\hat{\beta}_j) = c_{jj}\sigma^2/L_{jj}, j = 1, 2, \cdots, p$$

其中 L_{jj} 是 x_j 的离差平方和，用 c_{jj} 做为衡量自变量 x_j 的方差扩大程度的因子是恰如其分的。记 R_j^2 为自变量 x_j 对其余 $p-1$ 个自变量的复决定系数，可以证明

$$c_{jj} = \frac{1}{1 - R_j^2}$$

也可以作为放长扩大因子 VIF_j 的定义，由此式可知，$VIF_j \geq 1$

R_j^2 度量了自变量 x_j 与其余 $p-1$ 个自变量的线性相关程度，这种相关程度越强，说明自变量之间的多重共线性越严重，R_j^2 越接近于 1，VIF_j 就越大。

- 经验表明, 当 $VIF_j \geq 10$ 时, 就说明自变量 x_j 与其余自变量之间有严重的多重共线性, 且这种多重共线性可能会过度地影响最小二乘估计值。
- 还可用 p 个自变量所对应的方差扩大因子的平均数来度量多重共线性。
- 当

$$\overline{VIF} = \frac{1}{p}\sum_{j=1}^{p} VIF_j$$

远远大于 1 时就表示存在严重的多重共线性问题。

特征根判定法

- 根据矩阵行列式的性质，矩阵的行列式等于其特征根的连乘积。因而，当行列式 $|\boldsymbol{X}'\boldsymbol{X}| \approx 0$ 时，矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 至少有一个特征根近似为零。反之可以证明，当矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 至少有一个特征根近似为零时，$\boldsymbol{X}$ 的列向量间必存在复共线性，

证明:

记 $\boldsymbol{X} = (\boldsymbol{X}_0, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_p)$，其中 $\boldsymbol{X}_i, i = 0, 1, \cdots, p$ 为 $\boldsymbol{X}$ 的列向量，$\boldsymbol{X}_0 = (1, 1, \cdots, q)'$ 是元素全为 1 的 n 维列向量。λ 是矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 的一个近似为零的特征根，$\lambda \approx 0$，$\boldsymbol{c} = (c_0, c_1.c_2, \cdots, c_p)'$ 是对应于特征根 λ 的单位特征向量，则

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c} = \lambda\boldsymbol{c} \approx \boldsymbol{0}$$

上式两边左乘 $\boldsymbol{c}'$ 得 $\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c} \approx 0$

从而有 $\boldsymbol{X}\boldsymbol{c} \approx \boldsymbol{0}$ 即 $c_0\boldsymbol{X}_0 + c_1\boldsymbol{X}_1 + c_2\boldsymbol{X}_2 + \cdots + c_p\boldsymbol{X}_p \approx \boldsymbol{0}$

写成分量得形式

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} \approx 0$$

这正是定义的多重共线性关系。

- 如果矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 有多个特征根近似为零，在上面的证明中，取每个特征根的特征向量为标准化正交向量
- 证明：$\boldsymbol{X}'\boldsymbol{X}$ 有多少个特征根接近于零
- 设计矩阵 $\boldsymbol{X}$ 就有多少个多重共线性关系，并且这些多重共线性关系的系数向量就等于接近于零的那些特征根对应的特征向量。
- 特征根分析表明，当矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 有一个特征根近似为零时，设计矩阵 $\boldsymbol{X}$ 的列向量间必存在复共线性。那么特征根近似为零的标准如何确定？
- 记 $\boldsymbol{X}'\boldsymbol{X}$ 的最大特征根为 $\lambda_{\max}$，称

$$k_i = \sqrt{\frac{\lambda_{\max}}{\lambda_i}}, i = 0, 1, 2, \cdots, p$$

为特征根 λ_i 的条件数（Condition Index）。

- 用条件数判断多重共线性的准则
  - $0 < k < 10$ 时，设计矩阵 $\boldsymbol{X}$ 没有多重共线性；
  - $10 \le k < 100$ 时，认为 $\boldsymbol{X}$ 存在较强的多重共线性；
  - 当 $k \ge 100$ 时，则认为存在严重的多重共线性。

直观判定法

(1) 当增加或剔除一个自变量，其它自变量的系数估计值或显著性发生较大变化，则回归方程存在严重的多重共线性。

(2) 当定性分析认为重要的一些自变量在回归方程中没有通过显著性检验时，可初步判断存在着严重的多重共线性。

(3) 与因变量简单相关系数绝对值很大的自变量，在回归方程中没有通过显著性检验时，可初步判断存在着严重的多重共线性。

(4) 有些自变量的回归系数的数值大小与预期相差很大，甚至正负号与定性分析结果相反时，存在严重多重共线性问题。

(5) 自变量的相关矩阵中，自变量间的相关系数较大时，会出现多重共线性问题。

(6) 一些重要的自变量的回归系数的标准误差较大时，我们认为可能存在多重共线性。

1.3.3.4 消除多重共线性的方法

- 剔除一些不重要的解释变量
  - 在选择回归模型时, 可以将回归系数的显著性检验、方差扩大因子 VIF 的数值、以及自变量的经济含义结合起来考虑，以引进或剔除变量。
- 增大样本容量
  - 例如，我们的问题设计两个自变量 x_1 和 x_2，假设 x_1 和 x_2 都已经中心化。

$$var(\hat{\beta}_1) = \frac{\sigma^2}{(1-r_{12}^2)L_{11}}$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{(1-r_{12}^2)L_{22}}$$

    可以看到, 在 r_{12} 固定不变时，当样本容量 n 增大时,L_{11} 和 L_{22} 都会增大，两个方差均可减小，从而减弱了多重共线性对回归方程的影响。
- 回归系数的有偏估计
  - 岭回归法、主成分回归法、偏最小二乘法等。

1.4 变量选择

- 一个好的回归模型，并不是考虑的自变量越多越好。
- 在建立回归模型时，选择自变量的基本指导思想是“少而精”, 哪怕我们丢掉了一些对因变量 y 还有些影响的自变量，
- 由选模型估计的保留变量的回归系数的方差，要比由全模型所估计的相应变量的回归系数的方差小。
- 对于所预测的因变量的方差来说也是如此。丢掉了一些对因变量 y 有影响的自变量后，所付出的代价是估计量产生了有偏性。然而，尽管估计量是有偏的，但预测偏差的方差会下降。
- 如果保留下来的自变量有些对因变量无关紧要，那么，方程中包括这些变量会导致参数估计和预测的有偏性和精度降低。

1.4.1 所有子集回归

- $x_1, x_2, \cdots, x_m$
- 每个自变量都有入选和不入选两种情况，这样 y 关于这些自变量的所有可能的回归方程就有 $2^m - 1$ 个。这里减一是要求回归模型中至少包含一个自变量。
- 包含常数项

$$C_m^0 + C_m^1 + \cdots + C_m^m = 2^m$$

1.4.2 关于自变量选择的几个准则

- 从数据与模型拟合优劣的直观考虑出发，认为残差平方和 SSE 最小的回归方程就是最好的。
- 复相关系数 R^2 来衡量回归拟合的好坏。

然而这两种方法都有明显的不足，这是因为:

$$SSE_{p+1} \leq SSE_p$$

$$R_{p+1}^2 \geq R_p^2$$

准则 1 自由度调整复决定系数达到最大

- 调整的复决定系数为

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

- $R_a^2 \le R^2$，R_a^2 随着自变量的增加并不一定增大。因为尽管 $1-R^2$ 随着变量的增加而减小，但由于其前面的系数 $(n-1)/(n-p-1)$ 增大起折扣作用。
- 从拟合优度的角度追求最优，则所有回归子集中 R_a^2 最大者对应的回归方程就是最优方程。

从另一个角度考虑回归的拟合效果，

- 回归误差项方差 σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{1}{n-p-1}SSE$$

此无偏估计式中也加入了惩罚因子 $n-p-1$

- 残差平方和和复决定系数 R_a^2 有什么关系?

$$R_a^2 = 1 - \frac{n-1}{SST}\hat{\sigma}^2$$

由于 SST 是与回归无关的固定值，因此 R_a^2 与 $\hat{\sigma}^2$ 是等价的。

准则 2 AIC 与 BIC 准则

- AIC 准则是日本统计学家赤池 (Akaike)1974 年根据极大似然估计原理提出的一种较为一般的模型选择准则，人们称它为 Akaike 信息量准则 (Akaike Information Criterion, 简记为 AIC)。
- AIC 准则既可用来作回归方程自变量的选择，又可用于时间序列分析中自回归模型的定阶上。
- 由于该方法的广泛应用，使得赤池乃至日本统计学家在世界的声誉大增。

设模型的似然函数为 $L(\boldsymbol{\theta}, \boldsymbol{x})$，$\boldsymbol{x}$ 的维数为 p，为随机样本（在回归分析中随机样本为 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$），则 AIC 定义为:

$$AIC = -2\ln L(\hat{\boldsymbol{\theta}}_L, \boldsymbol{x}) + 2p \qquad (*)$$

其中 $\hat{\boldsymbol{\theta}}_L$ 为 $\boldsymbol{\theta}$ 的最大似然估计，p 为未知参数的个数。

假定回归模型的随机误差项 ϵ 服从正态分布，即

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

对数似然函数为

$$\ln L_{max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_L^2) - \frac{1}{2\hat{\sigma}_L^2}SSE$$

将 $\hat{\sigma}_L^2 = \frac{1}{n}SSE$ 代入得

$$\ln L_{max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\frac{SSE}{n}) - \frac{n}{2}$$

将上式代入 $(*)$，这里似然函数中未知参数得个数为 $P+2$，略去与 p 无关得常数，则回归模型得 AIC 公示为

$$AIC = n\ln(SSE) + 2p$$

对每一个回归子集计算 AIC，其中 AIC 最小者所对应得模型是最优回归模型。

- 赤池于 1976 年对 AIC 准则给予了改进，而施瓦茨 (Schwartz) 在 1978 年根据 Bayes 理论也得出同样的判别准则，称为 BIC 准则 (Bayesian information criterion)，也称为 SBC(Schwartz's Bayesian criterion) 准则，加大了对自变量数目的惩罚力度，
- BIC 达极小。

$$BIC = n\ln(SSE) + \ln(n)p \qquad (5.11)$$

- R 软件可以计算 BIC，计算形式大致为

$$BIC = n\ln\left(\frac{SSE}{SST}\right) + 1 + \ln(2\pi) + \ln(n)p \tag{5.12}$$

式 (5.11) 与 (5.12) 是等价的，两者的差值只与 n 和 SST 有关，与 p 无关。

准则 3 C_p 统计量达到最小

1964 年马勒斯 (Mallows) 从预测的角度提出一个可以用来选择自变量的统计量——C_p 统计量。根据性质 5，即使全模型正确，但仍有可能选模型有更小的预测误差。C_p 正是根据这一原理提出来的。

考虑在 n 个样本点上，用选模型式作回报预测时，预测值与期望值的相对偏差平方和为：

$$\begin{aligned} J_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_{ip} - E(y_i))^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{\beta}_{0p} + \hat{\beta}_{1p}x_{i1} + \cdots + \hat{\beta}_{pp}x_{ip} - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}))^2 \end{aligned}$$

J_p 的期望是

$$E(J_p) = \frac{E(SSE_p)}{\sigma^2} - n + 2(p+1)$$

略去无关的常数 2，据此构造出 C_p 统计量为

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p = (n-m-1)\frac{SSE_p}{SSE_m} - n + 2p$$

其中 $\hat{\sigma}^2 = \frac{1}{n-m-1}SSE_m$，为全模型中 σ^2 的无偏估计。

这样我们得到一个选择变量的 C_p 准则：选择使 C_p 最小的自变量子集，这个自变量子集对应的回归方程就是最优回归方法。

1.4.3 逐步回归

- 变量的所有可能子集构成 $2^m - 1$ 个回归方程，
- 当可供选择的自变量不太多时，用前边的方法可以求出一切可能的回归方程，然后用几个选元准则去挑出“最好”的方程，
- 但是当自变量的个数较多时，要求出所有可能的回归方程是非常困难的。

为此，人们提出了一些较为简便、实用、快速的选择“最优”方程的方法。人们所给出的方法各有优缺点，至今还没有绝对最优的方法，

- 目前常用的方法有“前进法”、“后退法”、“逐步回归法”，而逐步回归法最受推崇。
- 在后边的讨论中，无论我们从回归方程中剔除某个自变量，还是给回归方程增加某个自变量都要利用偏 F 检验，这个偏 F 检验与 t 检验是等价的，F 检验的定义式的统计意义更为明了，并且容易推广到对多个自变量的显著性检验，因而采用 F 检验。

$$F_j = \frac{\Delta SSR_{(j)}/1}{SSE/(n-p-1)}, \quad t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$$

1.4.3.1 前进法

前进法的思想是变量由少到多，每次增加一个，直至没有可引入的变量为止。具体的做法是首先将全部 m 个自变量分别对因变量 y 建立一元线性回归方程，并分别计算这 m 个一元线性回归方程的 m 个回归系数的 F 检验值，记为 $\{F_1^1, F_2^1, \cdots, F_m^1\}$，选其最大值记为

$$F_j^1 = \max\{F_1^1, F_2^1, \cdots, F_m^1\}$$

给定显著水平 α，若 $F_j^1 \geq F_\alpha(1, n-2)$，则首先将 x_j 引入回归方程，为了方便，设 x_j 就是 x_1。
接下来因变量 y 分别与 $(x_1, x_2), (x_1, x_3), \cdots, (x_1, x_m)$ 建立二元线性回归方程，对这 $m-1$ 个回归方程中 $x_2, \cdots, x_m$ 的回归系数进 F 检验，计算 F 值，记为 $\{F_2^2, F_3^2, \cdots, F_m^2\}$，选其最大值记为

$$F_j^2 = \max\{F_2^2, F_3^2, \cdots, F_m^2\}$$

若 $F_j^2 \geq F_\alpha(1, n-3)$，则将 x_j 引入回归方程。
依上述方法接着做下去。直至所有未被引入方程的自变量的 F 值均小于 $F_\alpha(1, n-p-1)$ 时为止。这时，得到的回归方程就是最终确定的方程。
每步检验中的临界值 $F_\alpha(1, n-p-1)$ 与自变量数目 p 有关，在用软件计算时，我们实际使用的是显著性 P 值（或记为 sig）做检验。例 5.4

1.4.3.2 后退法

- 后退法与前进法相反，首先用全部 m 个变量建立一个回归方程，然后在这 m 个变量中选择一个最不重要的变量，将它从方程中剔除。
- 设对 m 个回归系数进行 F 检验，记求得的 F 值为 $\{F_1^m, F_2^m, \cdots, F_m^m\}$，选其中最小者记为：

  $$F_j^m = \min\{F_1^m, F_2^m, \cdots, F_m^m\}$$

  给定显著水平 α，若 $F_j^m \leq F_\alpha(1, n-m-1)$，则首先将 x_j 从回归方程中剔除，为了方便，设 x_j 就是 x_m。

- 接着对剩下的 $m-1$ 个自变量重新建立回归方程，进行回归系数的显著性检验，像上面那样计算出 F_1^{m-1}，如果又有 $F_j^{m-1} \leq F_\alpha(1, n-(m-1)-1)$，则剔除 x_j，重新建立关于 $m-2$ 个自变量的回归方程，
- 以此类推，直至回归方程中所剩余的 p 个自变量的 F 检验值均大于临界值 $F_\alpha(1, n-p-1)$，没有可以剔除的变量为止。这时，得到的回归方程就是最终确定的方程。

1.4.4 逐步回归法

- 逐步回归的基本思想是“有进有出”。具体做法是将变量一个一个引入，当每引入一个自变量后，对已选入的变量要进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。
- 这个过程反复进行，直到既无显著的自变量选入回归方程，也无不显著自变量从回归方程中剔除为止。
- **优点**避免了前进法和后退法各自的缺陷，保证了最后所得的回归子集是“最优”回归子集。
- 在逐步回归中需要注意的一个问题是引入自变量和剔除自变量的显著性水平 α 值是不相同的，要求$\alpha_{进} < \alpha_{出}$，否则可能产生“死循环”。
  - 当 $\alpha_{进} \geq \alpha_{出}$ 时，
  - 某个自变量的显著性 P 值在 $\alpha_{进}$ 与 $\alpha_{出}$ 之间，那末这个自变量将被引入、剔除、再引入、再剔除、$\cdots$，循环往复，以至无穷.

第 2 章 变量选择

假设回归模型是

$$Y = X^\top \beta + \varepsilon \tag{2.1}$$

其中 Y 是一维随机变量，$X = (X_1, \ldots, X_p)$ 是 p 维随机变量，ε 是一维随机变量。$\beta = (\beta_1, \ldots, \beta_p)^\top$ 是未知参数。假设

$$E(\varepsilon | X = x) = 0, \tag{2.2}$$

模型 (2.1) 可以表示为：

$$E(Y | X = x) = x^\top \beta. \tag{2.3}$$

上述模型是均值回归 (Mean regression)，其参数可以通过下面得到:

$$\min E(Y - x\beta)^2 \tag{2.4}$$

备注 假条条件 (2.2) 不同，可以得到不同类别的统计模型，比如分位数回归 (Quantile regression) 和众数回归 (Mode regression)。这里主要讨论均值回归。

假设 $(\boldsymbol{x}_1, y_1)$,...,$(\boldsymbol{x}_n, y_n)$ 是一组样本，其中 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$，表达式 (2.4) 的样本实现值为

$$\min_\beta \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \beta)^2 \tag{2.5}$$

经过简单运算，$\hat{\beta}_{ols} = (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top y$，其中 $y = (y_1, \ldots, y_n)^\top$。$\boldsymbol{x} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$ 是 $n \times p$ 的设计矩阵 (Design matrix，矩阵是设计好的，换句话为是给定，直白点说就是已知的) 在条件 (2.2)，$\hat{\beta}_{ols}$ 是无偏估计（注意没有其它假设条件，如果想证明其它的性质，如渐近性质或，需要其它条件。

现在，如果 $\boldsymbol{x}^\top \boldsymbol{x}$ 不可逆，也就是说 $\boldsymbol{x}$ 不是列满秩，那么 $\hat{\beta}_{ols}$ 的不存在。上面这种线性成为完全共线性。这也是为什么研究变量选择的一个重要原因。下面我们来讨论另一个原因，假如研究儿童身高的影响因素，我们收集了性别、体重、父亲体重、母亲体重、家里花草的数量等几百个因素，目的是找到主要影响因素。大家注意，我们这里其实有一个假设，那就是儿童身高的影响因素是很少的，也只有几个。换句统计的词汇就是“稀疏性假设”。“家里花草的数量”显然就是需要排除的因素。排除因素就是变量选择。除了上述原因外，还有

- 估计量的方差变大，预测的精度较低；
- 过拟合，保留大量的解释变量会降低模型的可解释性。

怎么进行变量选择或者消去共线性，我们学习了很多方法，比如最有子集方法 (best subset method)，逐步回归和岭回归 (Ridge regression) 等。下面简单介绍这几种方法：

最优子集方法对 p 个变量的所有可能组合分别进行拟合，选择残差平方和 (Residual square sum) 或者 R^2 最小的模型。最优子集的优点是简单直观，但效率太低，当 p 很大时，从一个巨大的搜索空间中得到的模型通常会有过拟合和系数估计方差高的问题；改进的子集选择还有逐步选择 (向前、向后)，与全子集相比限制了搜索空间，提高了运算效率，但是无法保证找到的模型是 2^p 个模型中最优的。

岭回归是求带有约束的凸优化问题：

$$\min_\beta \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \beta)^2 \tag{2.6}$$

$$\text{s.t.} \sum_{j=1}^p \beta_j^2 \le t \tag{2.7}$$

其中 s.t. 是 subject to 的缩写，表示约束条件。$t > 0$。引入 Lagrange 乘子可以转化为，上面的优化问题可以转化为

$$\min_\beta \left(\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right) \tag{2.8}$$

岭回归得到的估计量是有偏的，但方差小了，得到的均方误差小，也就是其牺牲了无偏性，降低了方差。

2.1 LASSO 及其拓展

LASSO(Least absolute shrinkage and selection operator) 是[Tibshirani1996]提出，其求下面目标函数最小值

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \beta)^2 \tag{2.9}$$

$$\text{s.t.} \sum_{j=1}^{p} |\beta_j| \le t \tag{2.10}$$

上式等价于

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{2.11}$$

其中 λ 是截断参数 (Tuning parameter)。我们可以采用 ([Wang2007]) 方法，

$$\text{BIC}(\lambda_n) = \log \left(\sum_{i=1}^{n} (Y_i - Z_i^\top \beta) \right)^2 + \frac{\log n}{n} \times df$$

其中 df 估计非零的参数的个数。$\lambda_{opt} = \arg\min_{\lambda_n} \text{BIC}(\lambda_n)$.
相比岭回归，LASSO 只是将约束条件修改为绝对值。这样做为什么可以选择变量？从图 **??**，LASSO 更有可能得到稀疏的解，即某一个解为 0。这是由于解易出现菱角或者边缘。对于岭回归而言是约束域是圆，所以每一点的可能性相同，而矩阵有几个角，角的可能性更大些。

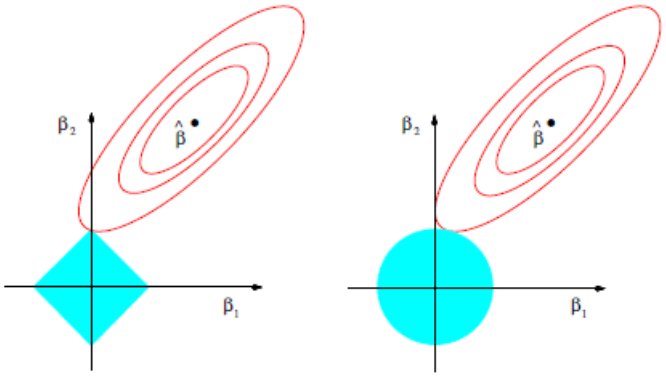

图 2.1: LAOO 和岭回归的几何解释；左边是 LASSO，右图是岭回归

LASSO 被提出后，后面有很多文章提出了不同的方法，如 SCAD ([Fan2001]) 和 Adaotive LASSO ([Zou2006]). 一般而言，后者更为简单，其为:

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right) \tag{2.12}$$

其中 $w_j = \dfrac{1}{|\widetilde{\beta}_j|^\kappa}$, and $\kappa > 0$. $\widetilde{\beta}$ 最小二乘的解[Zou2006]建议 $\kappa = 1$。
下面我们讨论几种变量选择的关系。假设 $\boldsymbol{x}^\top \boldsymbol{x} = I$，其中 I 是单位矩阵。所以 $\hat{\beta}_{ols} = \boldsymbol{x}^\top y, \hat{y}_{ols} = \boldsymbol{x}\boldsymbol{x}^\top y$，且 $\boldsymbol{x}^\top (y - \hat{y}_{ols}) = 0$。我们考虑一般的变量选择的惩罚函数形式分析:

$$\frac{1}{2} \|y - \boldsymbol{x}\beta\|^2 + \lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{2.13}$$

其中 $p_\lambda(\cdot)$ 是罚函数。经过计算，我们可以得到：

$$\begin{aligned}&\frac{1}{2}\|y-\boldsymbol{x}\boldsymbol{\beta}\|^2+\lambda\sum_{j=1}^{p}p_\lambda(|\beta_j|)\\&=\frac{1}{2}\|y-\hat{y}_{ols}|^2+\frac{1}{2}\sum_{j=1}^{p}\|\hat{\boldsymbol{\beta}}_{ols,j}-\boldsymbol{\beta}_j\|^2+\lambda\sum_{j=1}^{p}p_\lambda(|\beta_j|)\end{aligned}$$

这是因为:

$$\begin{aligned}&(y-\boldsymbol{x}\boldsymbol{\beta})^\top(y-\boldsymbol{x}\boldsymbol{\beta})\\&=(y-\hat{y}+\hat{y}-\boldsymbol{x}\boldsymbol{\beta})^\top(y-\hat{y}+\hat{y}-\boldsymbol{x}\boldsymbol{\beta})\\&=(y-\hat{y})^\top(y-\hat{y})+(\hat{y}-\boldsymbol{x}\boldsymbol{\beta})^\top(\hat{y}-\boldsymbol{x}\boldsymbol{\beta})+2(\hat{y}-\boldsymbol{x}\boldsymbol{\beta})^\top(y-\hat{y})\\&=\|y-\hat{y}\|^2+(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta})^\top\boldsymbol{x}^\top\boldsymbol{x}(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta})+2(\hat{y}-\boldsymbol{x}\boldsymbol{\beta})^\top(y-\hat{y})\end{aligned}$$

模型 (2.13) 可以转化为

$$\frac{1}{2}(\hat{\boldsymbol{\beta}}_{ols,j}-\boldsymbol{\beta}_j)^2+\lambda p_\lambda(|\boldsymbol{\beta}_j|),$$

更为一般的形式为：

$$\frac{1}{2}(z-\theta)^2+\lambda p_\lambda(|\theta|),$$

$p_\lambda(\cdot)$ 取不同的形式对应不同的方法：

1. $p_\lambda(\theta)=\lambda^2-(|\theta|-\lambda)I(|\theta|<\lambda)|\theta|^2$ 是最优子集估计量。
2. $p_\lambda(\theta)=\lambda|\theta|^2$ 是岭回归估计量
3. $p_\lambda(\theta)=\lambda|\theta|$ 是 LASSO 估计量
4. $p'_\lambda(\theta)=\lambda\Big\{I(\theta<\lambda)+\frac{(a\lambda-\theta)_+}{(a-1)\lambda}I(\theta\geq\lambda)$ 是 SCAD 估计量

经过推断，我们可以得到如下结论：

1. 最优子集的估计量 $\hat{\theta}=zI(|z|>\lambda)$
2. 岭回归估计量 $\hat{\theta}=\frac{z}{1+2\lambda}$
3. LASSO 估计量 $\hat{\theta}=\mathrm{sgn}(z)(|z|-\lambda)_+$
4. SCAD 估计量

$$\hat{\theta}=\begin{cases}\mathrm{sgn}(z)(|z|-\lambda)_+ & |z|\leq 2\lambda\\ \frac{(a-1)z-\mathrm{sgn}(z)a\lambda}{a-2} & 2\lambda\leq|z|\leq a\lambda\text{x!=0}\\ z & |z|\geq a\lambda\end{cases}$$

下面我们介绍[Fan2001]提出好的罚函数应该具备如下性质：

1. 无偏性 (Unbiasedness)：对于较大的 θ，$p'_\lambda(|\theta|)=0$，则 $\hat{\theta}=z$。
2. 稀疏性 (Sparsity)：$\min_{\theta\neq0}\{|\theta|+p'_\lambda(|\theta|)\}>0$，则解具有稀疏性，即当 $|z|<\min_{\theta\neq0}\{|\theta|+p'_\lambda(|\theta|)\}$ 时,$\hat{\theta}=0$.
3. 连续性 (Continuity)：$\min\{|\theta|+p'_\lambda(|\theta|)\}$

经过计算，我们可以得到

表 2.1: 几种罚函数的比较

方法	无偏性	稀疏性	连续性
最优子集	√	√	
岭回归			√
LASSO		√	
SCAD	√	√	√

2.2 组变量选择

顾名思义，组变量是一组变量。如分类变量具有 3 个水平，其需要转化为 2 个虚拟变量 (Dummy variable)。这 2 个虚拟变量是一组。我们进行变量选择，不能只选择其中的一个变量保留另一个变量。为了解决这个问题，[YuanLi2006]提出了组变量 LASSO，[HuangMa2012]详细总结了组变量选择方法 LASSO、SCAD 和 MCP，并且简单介绍了其在可加模型和

变系数模型的应用。假设 $(X_1, \ldots, X_p)$ 可以分成 K 组，其中每一组的自变量个数为 d_k，则一般表达式为：

$$\frac{1}{2n}\|y - \sum_{k=1}^{K} X_k\beta_k\|_2^2 + \sum_{k=1}^{K} p_\lambda(\|\beta_k\|_{R_k})$$

其中 $\|v\|_R^2 = v^\top Rv$. 通常 R 是一个单位矩阵。`grpreg` 包中的函数 `grpreg` 可以实现 LASS，SCAD 和 MCP 的组变量选。

```
grpreg(X, y, group=1:ncol(X), penalty=c("grLasso", "grMCP", "grSCAD"),
            family=c("gaussian", "binomial", "poisson"), .....
```

下面是利用 CV 准则选择 λ 的命令。

组变量选择代码

```
rm(list=ls())
library(grpreg)
data(Birthwt)
summary(Birthwt)
X <- Birthwt$X
y <- Birthwt$bwt
group <- Birthwt$group

cvfit <- cv.grpreg(X, y, group)
plot(cvfit)
summary(cvfit)
coef(cvfit) ## Beta at minimum CVE

```

参考文献

[Fan and Li(2001)] Fan J. and Li R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456): 1348–1360.

[Huang, Breheny & Ma (2012)] Huang, J., Breheny, P., & Ma, S. (2012). A selective review of group selection in high-dimensional models. Statistical science: a review journal of the Institute of Mathematical Statistics, 27(4).

[Lee, Kwon and Kin (2016)] Lee, S., Kwon, S. and Kim, Y. (2016). A modified local quadratic approximation algorithm for penalized optimization problems. Computational Statistics & Data Analysis, 94, 275-286.

[Yuan & Lin.(2006)] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49-67.

[Wang et al.(2007)] Wang H., Li R. and Tsai. C. (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 94(3): 553–568, 2007.

[Wang et al.(2009)] Wang H. Li B. and Leng C. (2009) Shrinkage tuning parameter selection with a diverging number of parameters, Journal of the Royal Statistical Society. Series B (Statistical Methodology), 71(3): 671–683.

[Tibshirani(1996)] Tibshirani R. (1996) Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 58(1): 267–288.

[Wu and Lang (2008)] Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. The Annals of Applied Statistics, 2(1), 224-244.

[Zhang and Huang(2008)] Zhang C. and Huang J.(2008) The sparsity and bias of the LASSO selection in highdimensional linear regression, The Annals of Statistics, 36(4): 1567–1594.

[Zou(2006)] Zou H.(2006) The adaptive LASSO and its oracle property. Journal of the American Statistical Association, 101(476): 1418–1429.

[Zou and Li(2008)] Zou H. and Li R. One-step sparse estimates in nonconcave penalized likelihood models. Annals of statistics, 36(4):1509–1533.

第 3 章 Generalized Linear Models

3.1 Introduction

3.1.1 Motivation Examples

Example 1: Mice Data (Binary). Twenty-six mice were given different amounts x_i of a drug. It was recorded whether they responded to the drug $(\boldsymbol{y_i = 1})$ or not $\boldsymbol{y_i = 0}$ We are concerned with how amount of the drug

Table 1: Mice Data: Dosage and Binary Response

$\boldsymbol{i}$	1	2	3	4	5	6	7	8	9	10	11	12	13
$\boldsymbol{x_i}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
$\boldsymbol{y_i}$	0	0	0	0	0	1	0	0	0	0	1	0	1
$\boldsymbol{i}$	14	15	16	17	18	19	20	21	22	23	24	25	26
$\boldsymbol{x_i}$	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5
$\boldsymbol{y_i}$	0	0	1	1	1	1	1	1	1	1	1	1	1

affects the responses of mice.

Example 2: Clinical Trial Data (Binomial Data). In a Phase I clinical trail to find the effective dose of a new drug, patients were randomly assigned to receive different doses of the drug. The table below shows the numbers m_i out of n_i patients who responded positively to the drug at each dose x_i.

Table 2: Clinical Trial Data: Binomial Responses

$\boldsymbol{i}$	1	2	3	4	5	6	7	8
$\boldsymbol{x_i}$	1.69	1.72	1.76	1.78	1.81	1.84	1.86	1.88
No. of patients, $\boldsymbol{n_i}$	59	60	62	56	63	59	62	60
No. responding, $\boldsymbol{m_i}$	6	13	18	28	52	53	61	60

The question is, how does the probability that a patient positively responds to the new drug depend on the baseline covariate $\boldsymbol{x_i}$?

Example 3: AIDS Data (Poisson count data). The table below gives the numbers of death from AIDS in Australia for three-month periods from 1983 to 1986.

Table 3: AIDS Data: Poisson count data

$\boldsymbol{i}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\boldsymbol{y_i}$	0	1	2	3	1	4	9	18	23	31	20	25	37	45

A scientific question is, how does the number of deaths from AIDS vary over time?

Remarks: These 3 examples highlight the difference from continuous data. This kind of data cannot be simply fitted using Normal linear models, as they are not normally distributed and the relationship between response and covariates may be non-linear. In fact, the data can be fitted well using generalised linear models.

3.2 Exponential family of distributions

3.2.1 Definition

Let $\boldsymbol{Y}$ be a random variable. If the probability density/mass function of $\boldsymbol{Y}$ has the form

$$f(y;\theta,\phi) = \exp\left\{\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

where

- θ is the so-called **natural parameter** of interest,
- ϕ is an additional scale or **nuisance/dispersion parameter,**

- $a(\cdot), b(\cdot)$ and $c(\cdot,\cdot)$ are known specific functions, then Y is said to have an **Exponential family of distributions** (EFD).

Example 4. Normal distribution. Let $Y \sim N\left(\mu, \sigma^2\right)$. Then the p.d.f of $\boldsymbol{Y}$ is

$$\begin{aligned} f(y;\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(y^2 - 2y\mu + \mu^2\right) - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right\} \\ &= \exp\left\{\left(y\mu - \frac{\mu^2}{2}\right)/\sigma^2 + \left(-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right)\right\} \end{aligned} \tag{2.2}$$

Comparing this to (2.1), we see

$$\begin{aligned} \theta &= \mu \text{ is the natural parameter} \\ a(\phi) &= \sigma^2 \text{ is the nuisance parameter} \\ b(\theta) &= \frac{\mu^2}{2} = \frac{\theta^2}{2} \text{ and } c(y,\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right) \end{aligned}$$

so that Normal distribution is a member of the EFD.

Example 5. Binomial distribution. Suppose we have a binary event (called a trial), i.e., the event has only two possible outcomes: "success" and "failure". Assume that the probability of success is $\boldsymbol{\pi}$. Let $\boldsymbol{Y}$ be the number of success in n independent trials. Then $\boldsymbol{Y}$ has a Binomial distribution with probability mass function

$$f(y;\pi) = \begin{pmatrix} n \\ y \end{pmatrix} \pi^y (1-\pi)^{n-y} \tag{2.3}$$

where $y = 0, 1, 2, \cdots, n$. We denote this as $\boldsymbol{Y \sim \mathcal{B}(n,\pi)}$. The p.m.f (2.3) can be further written into

$$\begin{aligned} f &= \exp\left\{y\log\pi + (n-y)\log(1-\pi) + \log\begin{pmatrix} n \\ y \end{pmatrix}\right\} \\ &= \exp\left\{y\log\frac{\pi}{1-\pi} + n\log(1-\pi) + \log\begin{pmatrix} n \\ y \end{pmatrix}\right\} \end{aligned} \tag{2.4}$$

Comparing to 2.1, we see that

$$\theta = \log\frac{\pi}{1-\pi}$$

so that $\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$ and $1-\pi = \frac{1}{1+\exp(\theta)}$. Hence, (2.4) can be written as

$$f(y;\pi) = \exp\left\{\left(y\theta - n\log\left(1+e^{\theta}\right)\right) + \log\begin{pmatrix} n \\ y \end{pmatrix}\right\} \tag{2.5}$$

In other words, we obtain that

$$\begin{aligned} \theta &= \log\frac{\pi}{1-\pi} \text{ is the natural parameter} \\ a(\phi) &= 1 \text{ is the nuisance parameter} \\ b(\theta) &= n\log\left(1+e^{\theta}\right) \\ c(y,\phi) &= \log\begin{pmatrix} n \\ y \end{pmatrix} \end{aligned}$$

so that Binomial distribution 2.3 is a member of the EFD.

Example 6: Bernoulli distribution. In the previous example, if $n = 1$ the Y is said to follow Bernoulli distribution, i.e., $Y \sim \mathcal{B}(1,\pi)$ or $Y \sim \mathcal{B}(\pi)$, and (2.5) becomes

$$f(y;\pi) = \exp\left\{\left(y\theta - n\log\left(1+e^{\theta}\right)\right) + \log\begin{pmatrix} 1 \\ y \end{pmatrix}\right\} \tag{2.6}$$

where $y = 0$ or 1 and $\log \begin{pmatrix} 1 \\ y \end{pmatrix} = \log \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \log 1 = 0$. Hence, we have

$$\begin{aligned} \theta &= \log \frac{\pi}{1-\pi} \text{ is the natural parameter} \\ a(\phi) &= 1 \text{ is the nuisance parameter} \\ b(\theta) &= n \log\left(1 + e^{\theta}\right) \\ c(y, \phi) &= 0 \end{aligned}$$

Again, it is a a member of the exponential family of distributions. (EFD)

Example 7: Poisson distribution. Poisson distribution is the distribution of the number of occurrences of some event in a defined time period or space, provided the occurrences of the event are independent. It is denoted as $\boldsymbol{Y \sim \mathcal{P}(\lambda)}$. The p.m.f of the Poisson distribution $\boldsymbol{\mathcal{P}(\lambda)}$ is

$$f(y; \lambda) = \frac{1}{y!} \lambda^{y} \exp(\lambda) \tag{2.7}$$

where $\boldsymbol{y = 0, 1, 2, \cdots}$

Note that (2.7) can be written as

$$f(y; \lambda) = \exp\{(y \log \lambda - \lambda) - \log y!\} \tag{2.8}$$

Comparing this to (2.1), we see that

$$\begin{aligned} \theta &= \log \lambda, \text{ is the natural parameter} \\ a(\phi) &= 1, \text{ is the nuisance parameter} \\ b(\theta) &= \lambda = \exp\{\theta\} \\ c(y, \phi) &= -\log y! \end{aligned}$$

Therefore, Poisson distribution belongs to the exponential family of distributions. (EFD)

We summarize the functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ for some commonly used distributions in the following table

Table 4: Summary of EFD

Distr,	θ	$a(\cdot)$	$b(\cdot)$	$c(\cdot, \cdot)$
$N\left(\mu, \sigma^2\right)$	μ	σ^2	$\frac{\mu^2}{2}$	$-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)$
$\mathcal{B}(n, \pi)$	$\log \frac{\pi}{1-\pi}$	1	$n \log\left(1 + e^{\theta}\right)$ $\left(= n \log \frac{1}{1-\pi}\right)$	$\log \begin{pmatrix} n \\ y \end{pmatrix}$
$\mathcal{B}(\pi)$	$\log \frac{\pi}{1-\pi}$	1	$n \log\left(1 + e^{\theta}\right)$ $(= n \log(1-\pi))$	0
$\mathcal{P}(\lambda)$	$\log \lambda$	1	$\exp\{\theta\} = \lambda$	$-\log y!$

3.2.2 Important properties

For the exponential family of distributions, we have the following important properties.

Property 1: For the exponential family of distribution in 2.1), the expectation and variance can be written as

$$E(Y) = b'(\theta), \quad \text{Var}(Y) = b''(\theta) a(\phi) \tag{2.9}$$

where $b'(\theta)$ and $b''(\theta)$ are the first- and second-derivatives of $b(\theta)$ with respect to θ

Proof: First, for any p.d.f $f(y; \theta, \phi)$, we have

$$\int_{-\infty}^{+\infty} f(y; \theta, \phi) dy = 1 \tag{2.10}$$

When taking derivative of (2.10) with respect to θ, we obtain

$$\frac{d}{d\theta} \int_{-\infty}^{+\infty} f(y; \theta, \phi) dy = 0.$$

Under certain regulations, it is equivalent to

$$\int_{-\infty}^{+\infty} \frac{d}{d\theta} f(y;\theta,\phi)dy = 0. \tag{2.11}$$

For the distribution in (2.1), since

$$\frac{d}{d\theta} f = f\left\{\frac{d}{d\theta}\log f\right\} \tag{2.12}$$

$$= f(y;\theta,\phi)\frac{d}{d\theta}\left\{\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right\} \tag{2.13}$$

$$= \frac{f(y;\theta,\phi)\,(y - b'(\theta))}{a(\phi)}$$

So that (2.11) becomes

$$\int_{-\infty}^{+\infty} \frac{(y - b'(\theta))}{a(\phi)} f(y;\theta,\phi)dy = 0$$

That is

$$\int_{-\infty}^{+\infty} y f(y;\theta,\phi)dy = \left(\int_{-\infty}^{+\infty} f(y;\theta,\phi)dy\right) b'(\theta)$$

i.e.,

$$\boldsymbol{E}(\boldsymbol{Y}) = b'(\theta)$$

Second, (2.11) also implies

$$\frac{d}{d\theta}\int_{-\infty}^{+\infty} \frac{df}{d\theta}(y;\theta,\phi)dy = \int_{-\infty}^{+\infty} \frac{d^2 f}{d\theta^2} f(y;\theta,\phi)dy = 0$$

While (2.12) gives

$$\frac{d^2 f}{d\theta^2} = \frac{df}{d\theta}\frac{(y - b'(\theta))}{a(\phi)} + f(y;\theta,\phi)\frac{-b''(\theta)}{a(\phi)}$$

$$= f(y;\theta,\phi)\frac{(y - b'(\theta))^2}{a^2(\phi)} + f(y;\theta,\phi)\frac{-b''(\theta)}{a(\phi)}$$

so that we obtain

$$0 = \frac{1}{a^2(\phi)}\int_{-\infty}^{+\infty} (y - b'(\theta))^2 f(y;,\theta,\phi)dy - \frac{b''(\theta)}{a(\phi)}\int_{-\infty}^{+\infty} f(y;,\theta,\phi)dy$$

$$= \frac{1}{a^2(\phi)}\mathrm{Var}(Y) - \frac{b''(\theta)}{a(\phi)}$$

or

$$\mathrm{Var}(Y) = b''(\theta)a(\phi)$$

and the proof is complete.

Note: Property 1, i.e., (2.9) provides a convenient way to calculate the mean and variance of a random variable which has a distribution from the exponential family of distributions.

Example 8: Normal distribution. From Example 2.1 we know that the functions $a(\phi)$ and $b(\theta)$ in $N\left(\mu,\sigma^2\right)$ are,

$$a(\phi) = \sigma^2 \text{ and } b(\theta) = \frac{\mu^2}{2}. \quad (\theta = \mu)$$

Therefore, we have

$$E(Y) = b'(\theta) = b'(\mu) = 2\mu/2 = \mu$$

$$\mathrm{Var}(Y) = b''(\theta)a(\phi) = b''(\mu)a(\phi) = 1\cdot\sigma^2 = \sigma^2.$$

which are the well-known results.

Example 9: Binomial distribution. For the Binomial distribution $\mathcal{B}(n,\pi)$, Example 2.1 shows that

$$a(\phi) = 1, b(\theta) = n\log(1 + \exp\{\theta\})$$

where

$$\theta = \log\frac{\pi}{1-\pi}$$

Therefore, we have

$$\begin{aligned} E(Y) = b'(\theta) &= n\frac{1}{1+\exp\{\theta\}}\exp\{\theta\} \\ &= n\frac{\pi/(1-\pi)}{1+\frac{\pi}{1-\pi}} = n\frac{\pi/(1-\pi)}{1/(1-\pi)} \\ &= n\pi \end{aligned}$$

i.e., $\boldsymbol{E(Y) = n\pi}$ and

$$\begin{aligned} \mathrm{Var}(Y) = b''(\theta)a(\phi) &= n\left(\frac{e^\theta}{1+e^\theta}\right)' \\ &= n\frac{e^\theta\left(1+e^\theta\right) - e^\theta e^\theta}{\left(e^\theta\right)^2} \\ &= n\frac{e^\theta}{1+e^\theta}\frac{1}{1+e^\theta} \\ &= n\frac{e^\theta}{1+e^\theta}\left(1-\frac{e^\theta}{1+e^\theta}\right) \\ &= n\pi(1-\pi) \end{aligned}$$

i.e., $\mathrm{Var}(Y) = n\pi(1-\pi)$.

Example 10: Bernoulli distribution. Bernoulli distribution is the Binomial distribution with $n = 1$ i.e., $Y \sim \mathcal{B}(1, \pi)$ or $Y \sim \mathcal{B}(\pi)$. Therefore,

$$E(Y) = \pi, \mathrm{Var}(Y) = \pi(1-\pi)$$

Example 11: Poisson distribution. For Poisson distribution $Y \sim \mathcal{P}(\lambda)$, Example 2.1 shows that

$$a(\phi) = 1, b(\theta) = \lambda = \exp\{\theta\}$$

so that

$$\begin{aligned} \boldsymbol{E(Y) = b'(\theta)} &= \exp\{\theta\} = \lambda, \\ \mathrm{Var}(\boldsymbol{Y}) = b''(\theta)a(\phi) &= \exp\{\theta\} = \lambda. \end{aligned}$$

That is, the expectation and variance are the same as λ, the occurrence rate of event.

The first derivative of $\log f(y;\theta,\phi)$ with respect to θ, called the **score function**, is equal to $U = \frac{\partial \log f}{\partial \theta}$. In addition, $-\frac{\partial U}{\partial \theta}\left(= -\frac{\partial^2 \log f}{\partial \theta^2}\right)$ and $E\left(-\frac{\partial^2 \log f}{\partial \theta^2}\right)$ are called the **observed information** and **Fisher information**, respectively.

For the exponential family of distribution (2.1), the log-density function is of the form,

$$\log f(y;\theta,\phi) = \frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)$$

so that

$$U = \frac{\partial \log f}{\partial \theta} = \frac{(y - b'(\theta))}{a(\phi)}$$

and

$$-\frac{\partial U}{\partial \theta} = -\frac{\partial^2 \log f}{\partial \theta^2} = \frac{b''(\theta)}{a(\phi)}$$

Furthermore,

$$\begin{aligned} E(U) &= \frac{E(Y) - b'(\theta)}{a(\phi)} = 0 \\ \mathrm{Var}(U) &= \frac{\mathrm{Var}(Y)}{a^2(\phi)} = \frac{a(\phi)b''(\theta)}{a^2(\phi)} = \frac{b''(\theta)}{a(\phi)} \\ -\frac{\partial U}{\partial \theta} &= \mathrm{Var}(U) \end{aligned}$$

$\Longrightarrow$

$$E\left(-\frac{\partial U}{\partial \theta}\right) = E\left(-\frac{\partial^2 \log f}{\partial \theta^2}\right) = E\left(\frac{\partial \log f}{\partial \theta}\right)^2$$

i.e.,

$$E\left(-\frac{\partial^2 \log f}{\partial \theta^2}\right) = E\left[\left(\frac{\partial \log f}{\partial \theta}\right)^2\right]$$

Property 2: For the EFD, we have

$$E\left(-\frac{\partial^2 \log f}{\partial \theta^2}\right) = E\left[\left(\frac{\partial \log f}{\partial \theta}\right)^2\right]$$

i.e., the Fisher information is equal to the expectation of the squared score function.

Property 2 indicates that when calculating the information we do not have to calculate the second-derivative of the log-density function with respect to the parameter. We only need to calculate the first-derivative and then take the expectation of the squared first-derivatives.

Note: Property 2 is important for the computation of the maximum likelihood estimate of $\boldsymbol{\theta}$.

3.3 Generalised linear models (GLMs)

3.3.1 Definition

The generalised linear models (GLMs) are defined by the following three components.

1. The **random components**:

The random samples $Y_1, Y_2, \cdots, Y_n$ comes from a distribution within the exponential family of distributions, that is, the distribution of $\boldsymbol{Y}_i$ is of the form

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

where the parameters of interest θ_i may vary with the index $i (i = 1, 2, \cdots, n)$, but the dispersion/nuisance parameter ϕ is a constant.

2. The **systematic components**:

For the i-th observation Y_i, we have a systematic component called **linear predictor**, which is a linear combination of some covariates, that is

$$\eta_i = x_i^T \beta = \sum_{j=1}^{p} x_{ij}\beta_j, i = 1, 2, \cdots, n \tag{3.1}$$

3. The **link function**:

There is a monotone and differentiable function $g(\cdot)$ called the link function, which links the expectation of random components and the systematic components through

$$g\left(\mu_i\right) = \eta_i = x_i^T \beta, i = 1, 2, \cdots, n$$

where $\boldsymbol{\mu}_i = \boldsymbol{E}\left(\boldsymbol{Y}_i\right)$ is the expectation of $\boldsymbol{Y}_i$.

Note: Denote

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}_{n\times 1}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}_{n\times 1}, \quad \text{and} \quad X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}_{n\times p} \tag{3.1}$$

then the link function can be expressed in matrix form

$$g(\mu) = \eta = X\beta \tag{3.2}$$

where $\boldsymbol{\beta}$ is the regression parameter vector of interest in GLMs.

Example 12: Normal linear model. The classical linear models with a Normal/Gaussian distribution is a special case of the GLMs. In fact, if the random samples

$$Y_1, Y_2, \cdots, Y_n \sim N\left(\mu_i, \sigma^2\right)$$

with $\mu_i = x_i^T \beta$, we know

1. The Normal distribution is a special member of the exponential family of distributions.
2. The systematic components are

$$\boldsymbol{\eta}_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

3. The identity link function:

$$g(\mu_i) = \mu_i = \eta_i = x_i^T \beta$$

In other words, the Normal linear model is the generalised linear model that has **Normal distribution** and an **identity link function.**

Example 13: Bernoulli-logistic model. Example 1 gives the dosage of a drug and the binary response of 26 mice in an experiment. The main concern is how the drug affects the probability of response of mice. In this example, we know

1. The random variable $Y_i \sim \mathcal{B}(\pi_i), \quad i = 1, 2, \cdots, 26$, i.e., the Bernoulli distribution is a special member of the exponential family of distributions.
2. The systematic component is

$$\eta_i = \beta_1 + x_i\beta_2 = (1, x_i)\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \equiv x_i^T\beta, i = 1, 2, \cdots, 26$$

where $x_i^T = (1, x_i), \beta = (\beta_1, \beta_2)^T$ in which β_1 is the intercept and $\boldsymbol{\beta_2}$ is the slope.

3. The expectation $\boldsymbol{\mu_i} = \boldsymbol{E}(\boldsymbol{Y}_i) = \pi_i > 0$ (Note that π_i is a probability while the linear predictor $\boldsymbol{\eta_i}$ might be negative). Thus we need a link function which links the probability π_i and the linear predictor $\boldsymbol{\eta_i}$.A natural way is to choose the natural parameter (see Table 4)

$$\log \frac{\pi_i}{1-\pi_i} = \eta_i = x_i^T\beta$$

i.e., $\boldsymbol{g}(\boldsymbol{\pi_i}) = \log \frac{\pi_i}{1-\pi_i}$, which is called logistic link function.

Note: When β is estimated, we then know how the dosage x_i affects the probability that the mice response positively to the drug.

Example 14: Binomial-logistic model. In Example 2, the numbers m_i out of n_i patients who responded positively to the drug at dosage x_i were recorded (see Table 2). We would model how the probability that a patient responds positively is related to the dosage $\boldsymbol{x_i}$.

1. The number of responses, random variable $Y_i \sim \mathcal{B}(n_i, \pi_i), i = 1, 2, \cdots, 8$, i.e., the Binomial distribution is a special member of the exponential family of distributions.
2. The systematic component is

$$\eta_i = x_i^T\beta = (1, x_i)\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} i = 1, 2, \cdots, 8$$

where $\boldsymbol{x_i}$ is the dosage of the drug taken by the patients in $\boldsymbol{i}$-th experiment, $\boldsymbol{\beta_1}$ is the intercept and $\boldsymbol{\beta_2}$ is the slope.

3. For the similar reason in Example 1.1.1, we need a link function which links $\boldsymbol{\pi}_i$, the probability of responding positively at dosage $\boldsymbol{x}_i$, and the baseline covariate $\boldsymbol{x_i}$. A natural way is to choose the natural parameter (see Table 4)

$$\log \frac{\pi_i}{1-\pi_i} = \eta_i = x_i^T\beta$$

i.e., $g(\pi_i) = \log \frac{\pi_i}{1-\pi_i}$, the logistic link function.

Example 15: Poisson-log model. In Example 3, the numbers of death from AIDS in Australia for three-month periods from 1983 to 1986 were recorded (see Table 3). We would model how the numbers of such death from AIDS, in average, vary over the time.

1. Let random variable $\boldsymbol{Y}_i$ be the number of patients who dead from AIDS in the i-th observation time, then

$$\boldsymbol{Y_i} \sim \mathcal{P}(\boldsymbol{\lambda_i})$$

where $\boldsymbol{\lambda_i}$ is the average number of death at $\boldsymbol{i}$-th observation. Obviously, it is is a special member of the exponential family of distributions.

2. The systematic component is

$$\eta_i = (1, t_i)\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = x_i^T\beta, (i = 1, 2, \cdots, 14)$$

where $\boldsymbol{t_i}$ is the time at which the number $\boldsymbol{Y_i}$ of death from AIDS was made.

3. The natural link for Poisson distribution is the logrithm function (see Table 4), i.e.,

$$\log \lambda_i = \eta_i = x_i^T\beta = \beta_1 + \beta_2 t$$

3.3.2 Link functions

1. As indicated in the definition of GLMs, a link function that is monotone and differentiable is used to link the expectation of random components and the systematic components.
2. Examples 12-15 choose the natural parameters associated with the distribution as the link functions. In this case, it is called **canonical link.** In other words, if the link function $\boldsymbol{g}(\cdot)$ takes the same form as the natural parameter, then it is called **canonical link function.**
3. The advantage of a canonical link function is that it leads to desirable statistical properties of GLMs and it is convenient to use. For example, for the most commonly used distribution, we have the following canonical links

Table 5: Canonical link functions

Normal	$\boldsymbol{\mu} = \boldsymbol{\eta}$, (identity link)
Poisson	$\log \boldsymbol{\mu} = \boldsymbol{\eta}$, (log link)
Bernoulli	$\log \frac{\pi}{1-\pi} = \boldsymbol{\eta}$, (logistic link)
Binomial	$\log \frac{\pi}{1-\pi} = \boldsymbol{\eta}$, (logistic link)

4. However, canonical link is not an unique choice. Others appropriate link functions in GLMs may include

(a) Probit link: $\eta = \Phi^{-1}(\pi), 0 < \pi < 1$ where $\Phi(\cdot)$ is the Normal cumulative distribution function (c.d.f).

(b) Complementary log-log link: $\eta = \log\{-\log(1-\pi)\}, 0 < \pi < 1$

(c) Power family of links:

$$\eta = \begin{cases} \mu^{\lambda}, & \text{if } \lambda =\neq 0 \\ \log \mu, & \text{if } \lambda = 0 \end{cases}$$

We will see the similarities and differences of these link functions when comparing to the logistic link in later sections.

3.4 实例分析

`plasma` 包含两个自变量 fibrinogen 和 globulin。目的研究两个自变量对 ESR 的影响。

图**??**的代码 1

```
> rm(list=ls())
> library(HSAUR3)
> data(plasma)
> head(plasma)
  fibrinogen globulin      ESR
1       2.52       38 ESR < 20
2       2.56       31 ESR < 20
3       2.19       33 ESR < 20
4       2.18       31 ESR < 20
5       3.41       37 ESR < 20
6       2.46       36 ESR < 20
```

```
> fit1 <- glm(ESR~fibrinogen+globulin, family = binomial,data = plasma)
> summary(fit1)

Call:
glm(formula = ESR ~ fibrinogen + globulin, family = binomial,
    data = plasma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9683  -0.6122  -0.3458  -0.2116   2.2636

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.7921     5.7963  -2.207   0.0273 *
fibrinogen    1.9104     0.9710   1.967   0.0491 *
```

```
globulin      0.1558     0.1195   1.303   0.1925
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ° 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 22.971  on 29  degrees of freedom
AIC: 28.971

Number of Fisher Scoring iterations: 5
```

我们发现 globulin 的 P 值大于 0.05，说明在显著水平为 0.05 的情况下，globulin 不显著。故我们删去改变量。

```
> fit2 <- glm(ESR~fibrinogen, family = binomial,data = plasma)
> summary(fit2)

Call:
glm(formula = ESR ~ fibrinogen, family = binomial, data = plasma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9298  -0.5399  -0.4382  -0.3356   2.4794

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.8451     2.7703  -2.471   0.0135 *
fibrinogen    1.8271     0.9009   2.028   0.0425 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ° 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 24.840  on 30  degrees of freedom
AIC: 28.84

Number of Fisher Scoring iterations: 5
```

删去 globulin 变量显著了。AIC 也降低了。