回归分析十五讲义

Regression Analysis Lecture Notes

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

1 介绍 **1**

1.1 变量之间的关系 2

1.2 实际问题 3

1.3 回归分析的发展 10

第 1 讲 练习 10

2 一元线性回归 **12**

2.1 引言 12

2.2 一元线性回归模型 21

2.3 最小二乘估计的性质 29

第 2 讲 练习 41

3 多元线性回归 **42**

3.1 diabetes 数据 42
3.2 多元线性回归模型的一般形式 43
3.3 回归参数的估计 47
3.4 约束的最小二乘 59
3.5 模型评价 61
3.6 残差分析 62
3.7 中心化和标准化 74
3.8 相关阵与偏相关系数 76
3.9 diabetes 数据分析 79
3.10 小结和评注 82
3.11 附录 A 83
3.12 附录 B 86
3.13 参考文献 88
第 3 讲 练习 88

4 违背基本假设的处理 **92**
4.1 引言 92

4.2 异方差性 93
4.3 自相关性问题及其处理 114
4.4 BOX-COX 变换 122
4.5 异常值与强影响点 124
4.6 参考文献 128
第 4 讲 练习 128

5 多重共线性的情形及其处理 135
5.1 diabetes 数据续 135
5.2 多重共线性 136
5.3 消除多重共线性的方法 147
5.4 岭回归 148
5.5 自变量选择的准备 151
5.6 所有子集回归 160
5.7 逐步回归 160
第 5 讲 练习 164

6 变量选择 165
6.1 LASSO 及其拓展 . . . 168
6.2 算法 . . . 174
6.3 组变量选择 . . . 187

7 广义线性模型 191
7.1 指数分布族 . . . 191
7.2 Generalised linear models (GLMs) . . . 207
7.3 Estimation procedure in GLMs . . . 215
7.4 Statistical Inferences . . . 234

8 多分类 Logit 模型 251
8.1 对数线性模型 . . . 251
8.2 多分类 Logit 模型 . . . 254
8.3 次序 Logsit 回归 . . . 263
8.4 连续比 logit 模型 . . . 268
8.5 实际数据分析 . . . 270

8.6 参考文献 273
第 8 讲 练习 275

9 分位数回归 276
9.1 总体分位数定义 276
9.2 分位数回归 280
9.3 参考文献 295
9.4 附录 297

10 众数回归 303
10.1 模型表达 304
10.2 MEM 算法 305
10.3 h 的选择 306
10.4 模拟实验 308
10.5 参考文献 313

11 固定效应和随机效应模型 314

11.1 引言 314
11.2 模型介绍和 R 实现 318

12 纵向数据回归 **327**

13 分层数据回归 **328**

14 超高维变量筛选 **329**

15 海量数据分析 **330**
15.1 线性回归 330

第 1 讲 介绍

- Galton(1822-1911) 在 1886 年发表了关于回归的开山论文《遗传结构中向中心的回归 (Regression towards mediocrity in heredity structure)》到现在是 130 多年
- 研究父代身高与子代身高之间的关系，发现：

$$\hat{y} = 33.73 + 0.156x$$

  - 子代平均身高介于其父代的身高和种群平均身高之间
  - 高个子父亲的儿子的身高低于其父亲身高的趋势
  - 矮个子父亲的儿子的身高则高于其父亲身高的趋势
  - 子代身高有向种群身高“回归”的趋势
- 现代的回归：
  - 探索和检验自变量 (X) 和因变量 (Y) 之间的关系：因果关系，数量关系
  - 基于自变量的取值变化预测因变量的取值
  - 描述自变量与因变量之间的关系

1.1 变量之间的关系

1.1.1 确定关系

函数关系

- 商品的销售额与销售量之间的关系

$$y = px$$

- 圆的面积与半径之间的关系

$$S = \pi R^2$$

- 原材料消耗额与产量 ($x1$) 、单位产量消耗 ($x2$) 、原材料价格 ($x3$) 之间的关系

$$y = x_1 x_2 x_3$$

1.1.2 相关关系

- 子女身高 (y) 与父亲身高 (x) 之间的关系
- 收入水平 (y) 与受教育程度 (x) 之间的关系

- 粮食亩产量 (y) 与施肥量 ($x1$) 、降雨量 ($x2$) 、温度 ($x3$) 之间的关系
- 商品的消费量 (y) 与居民收入 (x) 之间的关系
- 商品销售额 (y) 与广告费支出 (x) 之间的关系

图1.1 R 代码

```
rm(list=ls())
n <- 100
x <- rnorm(n)
y <- 1+ 2 * x + 0.5 * rnorm(n)
plot(x, y, pch=17)
lines(x, 1+ 2*x, col="red")
```

1.2 实际问题

1.2.1 NewYorkRivers 数据

- 数据名称：NewYorkRivers.txt

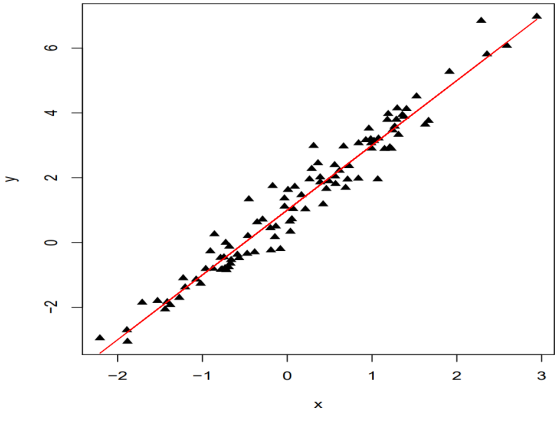


图 1.1: 模拟例子

- 变量说明：纽约州 20 条河流
  - Y:Nitrogen 平均氮浓度（毫升/升）
  - X_1:Agr 农业用地面积百分比
  - X_2:Forest 森林占地面积百分比
  - X_3:Rsdntial 住宅用地面积百分比
  - X_4:ComIndl 工业用地面积百分比
- 分析目的：河流流域土地利用状况对水质污染的影响

NewYorkRivers.txt

```
> rm(list=ls())
> dat_ryr <- read.table("NewYorkRivers.txt", head=TRUE)
> head(dat_ryr)
       River Agr Forest Rsdntial ComIndl Nitrogen
1      Olean  26     63      1.2    0.29     1.10
2  Cassadaga  29     57      0.7    0.09     1.01
3      Oatka  54     26      1.8    0.58     1.90
4  Neversink   2     84      1.9    1.98     1.00
```

5	Hackensack	3	27	29.4	3.11	1.99
6	Wappinger	19	61	3.4	0.56	1.42

1.2.2 EgyptianSkulls 数据

- 数据名称：EgyptianSkulls.txt
- 变量说明：
  - Y: 大概的年份（负值 = 公元前；正值 = 公元后）
  - X_1:MB 头盖骨的最大宽度
  - X_2:BH 头盖骨的颅最高点的高度
  - X_3:BL 头盖骨颅底牙槽的长度
  - X_4:NH 头盖骨的鼻高度
- 分析目的：推断埃及头骨的年代

EgyptianSkulls.txt

```
> rm(list=ls())
> dat_es <- read.table("EgyptianSkulls.txt", head=TRUE)
> head(dat_es)
```

```
    Year MB  BH  BL NH
1 -4000 131 138  89 49
2 -4000 125 131  92 48
3 -4000 131 132  99 50
4 -4000 119 132  96 44
5 -4000 136 143 100 54
6 -4000 138 137  89 56
```

1.2.3 Financial Ratios 数据

- 数据名称：Financial Ratios.txt
- 66 家企业的营运财务比，其中 33 家破产、33 家企业经营稳定
  - Y: 是否破产
  - X_1: 未支付利润/总资产
  - X_2: 支付利息和税金前利润/总资产
  - X_3: 销售额/总资产
- 分析目的：破产因素

Financial Ratios.txt

```
> rm(list=ls())
> dat_fr <- read.table("FinancialRatios.txt", head=TRUE)
> head(dat_fr)
  Y     X1     X2  X3
1 0  -62.8  -89.5 1.7
2 0    3.3   -3.5 1.1
3 0 -120.8 -103.2 2.5
4 0  -18.1  -28.8 1.1
5 0   -3.8  -50.6 0.9
6 0  -61.2  -56.2 1.7
```

1.2.4 NewDrugs 数据

- 数据名称：NewDrugs.txt
- 1992-1995 年 16 年中疾病引入的新药
  - D: 新药个个数

- P: 每十万人的病人数
- M:1994 年研究经费（百万美元）

- 分析目的：新药功效

```
> rm(list=ls())
> dat_nd <- read.table("NewDrugs.txt", head=TRUE)
> head(dat_nd)
                Disease  D     P      M
1  IschemicHeartDisease  6  8976  198.4
2            LungCancer  3   874   80.2
3               HIV/AIDS 21  1303 1049.6
4            AlcoholUse  2 18092  222.6
5 CerebrovascularDisease  2  9467  108.5
6                  COPD  1  4271   48.9
```

1.3 回归分析的发展

- 线性回归
- 分位数回归 (Quantile regression)
- 众数回归 (Mode regression)
- ……
- LASSO

第 1 讲 练习

1. 设 $y_1, \ldots, y_n$ 是一组样本，其中 μ 和 σ 都是未知的。构建下面模型

$$y_i = \mu + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

我们可以采用

- 最小二乘法估计: 最小化 $\sum_{i=1}^{2}(y_i - \mu)^2$
- 最小绝对值估计: 最小化 $\sum_{i=1}^{2}|y_i - \mu|$

回答下面问题:

(a). 证明 μ 的最小二乘估计是样本均值

(b). 证明 μ 的最小绝对值估计是样本中位数

(c). 列出样本均值的一个优点和一个缺点

(d). 列出样本中位数的一个优点和一个缺点

(e). 你会选择 μ 的两个估计量的哪一个？说出理由。

第 2 讲 一元线性回归

在我们讲解一元线性回归前，我们首先考虑数学期望和条件数学期望。

2.1 引言

通常

- $F(x) = P(X \leq x); -\infty < x < \infty$
- $f_X(x) = f(x) = F'(x)$
-

$$\mathbf{E}(X) = \sum_x x p(x)$$

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- $E(x)$ is the "center" of a distribution (or its r.v.) in the sense that

$$\min_b \mathbf{E}(X-b)^2 = \mathbf{E}[X-\mathbf{E}X]^2$$

- Conditional Expectation of X when Y is given as y is that
  - $\mathbf{E}(X|Y=y) = \sum_x x p_{X|Y}(x|y)$ for discrete r.v.
  - $\mathbf{E}(X|Y=y) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)dx$ for cont. r.v.
  - Interpretation: Note that $X|Y=y$ is a new r.v., $\mathbf{E}(X|Y=y)$ is the expectation on this r.v.
- Law of Total Expectation [1]

$$\mathbf{E}[\mathbf{E}(X|Y)] = \mathbf{E}(X)$$

- Law of Total Variance

$$Var(X) = Var[\mathbf{E}(X|Y)] + \mathbf{E}[Var(X|Y)]$$

[1] Statistical Inference 2nd Edition by George Casella Roger L. Berger

定义 2.1 (双期望公式)

If X and Y are any two r.v.s, then

$$\mathbf{E}(X) = \mathbf{E}\Big[\mathbf{E}(X|Y)\Big]$$

♣

Proof:

$$\begin{aligned}\mathbf{E}X &= \int\int x f(x,y)dxdy \\ &= \int \Big[\int x f(x|y)dx\Big] f_Y(y)dy \\ &= \int \mathbf{E}(X|y) f_Y(y)dy = \mathbf{E}\Big[\mathbf{E}(X|Y)\Big]\end{aligned}$$

In general, the conditional expectation $\mathbf{E}[X|Y]$ can by defined as a r.v. $g(Y)$ such that

$$\mathbf{E}[(X-g(Y))^2] = \inf_{among\ \ all\ \ reasonable\ \ function\ \ h} \mathbf{E}[(X-h(Y))^2]$$

or $\mathbf{E}[X|Y]$ is the function of Y which is "closest" to X in terms of mean square error.

定义 2.2 (双期望公式)

For any two random variables X and Y

$$Var(X) = Var[\mathbf{E}(X|Y)] + \mathbf{E}[Var(X|Y)]$$

provided that the expectation exist.

♣

$$\begin{aligned} Var(X) &= \mathbf{E}X\Big\{[X - \mathbf{E}(X|Y) + \mathbf{E}(X|Y) - \mathbf{E}]^2\Big\} \\ &= \mathbf{E}\Big\{[X - \mathbf{E}(X|Y)]^2 + [\mathbf{E}(X|Y) - \mathbf{E}X]^2 \\ &+ 2[X - \mathbf{E}(X|Y)][\mathbf{E}(X|Y) - \mathbf{E}X]\Big\} \\ &= \mathbf{E}\{[X - \mathbf{E}(X|Y)]^2\} + \mathbf{E}\{[\mathbf{E}(X|Y) - \mathbf{E}X]^2\} \\ &= \mathbf{E}[Var(X|Y)] + Var[\mathbf{E}(X|Y)] \end{aligned}$$

$$\mathbf{E}\Big\{2[X - \mathbf{E}(X|Y)][\mathbf{E}(X|Y) - \mathbf{E}X]\Big\} = \mathbf{E}[\mathbf{E}(Z|Y)]$$

例题 2.1

- The Current Population Survey (CPS) is a monthly survey of about 57,000 U.S. households conducted by the Bureau of the Census of the Bureau of Labor Statistics [cps09mar][2].
- female: 1 if female, 0 otherwise
- earnings: total annual wage and salary earnings
- hours: number of hours worked per week
- week: number of weeks worked per year

```
> rm(list=ls())
> dat20 <- read.table("cps09mar.txt", head=TRUE, fileEncoding="utf8")
> head(dat20)
 age female hisp education earnings hours week union uncov region race marital
1 52      0    0        12   146000    45   52     0     0      1    1       1
2 38      0    0        18    50000    45   52     0     0      1    1       1
3 38      0    0        14    32000    40   51     0     0      1    1       1
4 41      1    0        13    47000    40   52     0     0      1    1       1
5 42      0    0        13   161525    50   52     1     0      1    1       1
6 66      1    0        13    33000    40   52     0     0      1    1       1
```

- $X = \frac{earnings}{hours * week}$

[2]https://www.ssc.wisc.edu/ bhansen/econometrics/

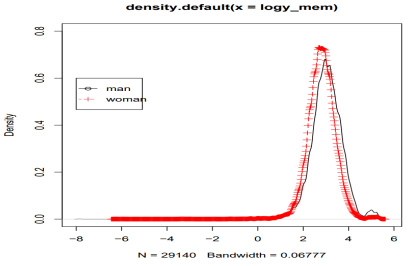


Figure: The density of $Y|female$

图 2.1: 核密度估计

- $Y = female$
- Homework: **Write R code to check the following equations.**

$$\mathbf{E}\Big[\mathbf{E}(X|Y)\Big] = \mathbf{E}(X)$$

$$Var(X) = Var[\mathbf{E}(X|Y)] + \mathbf{E}[Var(X|Y)]$$

cps09mar

```
1 y <- dat20$earnings/(dat20$hours*dat20$week)
```

```
logy <- log(y)
plot(density(log(y)))
index_men <- which(dat20$female==0)
index_women <- which(dat20$female==1)
logy_mem <- log(y[index_men])
logy_women <- log(y[-index_men])
plot(density(logy_mem), ylim=c(0, 0.8), pch=1)
points(density(logy_women), col="red", lty=2, pch=3)
legend(-8, 0.6, c("man", "woman"), col = c(1,2), lty = c(1, 2),
       pch = c(1, 3))
```

```
> mean(logy)
[1] 2.946185
> var(logy)
[1] 0.456827
> pp <- c((1-mean(dat20$female)), mean(dat20$female))
> mean_menwomean <- c( mean(logy_mem), mean(logy_women))
```

```
> var_menwomean <- c( var(logy_mem), var(logy_women))
> resu <- data.frame(pp, mean_menwomean, var_menwomean)
> rownames(resu) <- c("men","women")
> resu
            pp  mean_menwomean  var_menwomean
men   0.5742777        3.045938      0.4956187
women 0.4257223        2.811624      0.3729886
```

- 分析 female 对 Y 的影响
- 如果 female 是连续变量

R

```
y <- seq(-4,4,0.1)
x <- dnorm(y)
plot((x+0.1),y,xlim=c(0,4.5), ylim=c(-4,10),xlab="",
type="l", ylab="", xaxt="n", yaxt="n")
abline(v=min(x+0.1), col=3)
points((x+1),(y+1),type="l")
```

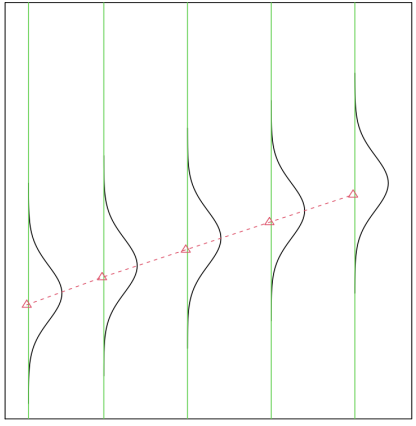


图 2.2: Caption

```
abline(v=min(x+1), col=3)
points((x+2),(y+2),type="l")
abline(v=min(x+2), col=3)
points((x+3),(y+3),type="l")
abline(v=min(x+3), col=3)
points((x+4),(y+4),type="l")
abline(v=min(x+4), col=3)
index <- which(x==max(x))
meanx <- x[index] + c(0.1, seq(1:4))
meany <- y[index] + c(0, seq(1:4))
points(meanx-0.42 ,meany-0.42, pch=2, col=2, type="o", lty=2)
```

2.2 一元线性回归模型

一元线性回归模型的数学形式为：

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

通常假定：

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

对式 (1) 两端求条件期望，得到回归方程：

$$E(y|x) = \beta_0 + \beta_1 x$$

如果获得 n 组样本观测值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，样本模型：

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \qquad i = 1, 2, \ldots, n \tag{2}$$

满足：

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

对 (2) 两端分别求期望和方差，得到：

$$E(y_i) = \beta_0 + \beta_1 x_i, \quad var(y_i) = \sigma^2, \quad i = 1, 2, \ldots, n$$

$E(y_i) = \beta_0 + \beta_1 x_i$ 从平均意义上表达了变量 y 与 x 的统计规律性。

用 $\hat{\beta}_0, \hat{\beta}_1$ 分别表示 β_0, β_1 的估计值，获得 y 关于 x 的一元线性经验回归方程

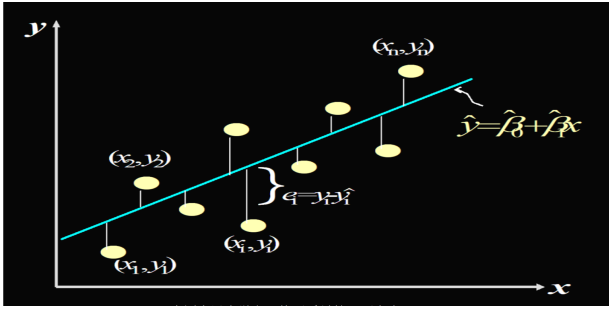
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

2.2.1 普通最小二乘估计

普通最小二乘估计 (Ordinary Least Square Estimation, 简记为 OLSE) 就是寻找参数 β_0, β_1 的估计值使离差平方和达到极小

$$(\hat{\beta}_0, \hat{\beta}_1)^\top = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- $\hat{\beta}_0, \hat{\beta}_1$ 称为 β_0, β_1 的最小二乘估计
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 为 y_i 的回归拟合值，简称回归值或拟合值
- $e_i = y_i - \hat{y}_i$ 为 y_i 的残差。
- 残差平方和 $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- 整体上刻画了 n 个样本观测点 $(x_i, y_i), i = 1, \ldots, n$ 到回归直线 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 距离的长短。



$$\begin{cases} \left.\dfrac{\partial Q}{\partial \beta_0}\right| = -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0 \\ \left.\dfrac{\partial Q}{\partial \beta_1}\right| = -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i=0 \end{cases}$$

可以得到残差的性质：$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases}$

整理后得到正规方程组

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

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases}$$

其中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$。

记：

$$L_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

$$L_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n(\bar{x}\bar{y})$$

则：

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} \end{cases}$$

例题 2.2

$$\bar{x} = \frac{49.2}{15} = 3.28, \quad \bar{y} = \frac{396.2}{15} = 26.413$$

$$L_{xx} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 196.16 - 15 \times (3.28)^2 = 34.784$$

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 1470.65 - 1299.536 = 171.114$$

得到：

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 26.413 - 4.919 \times 3.28 = 10.279 \\ \hat{\beta}_1 = L_{xy}/L_{xx} = 171.114/34.784 = 4.919 \end{cases}$$

于是得到回归方程：

$$\hat{y} = 10.275 + 4.919x$$

2.2.2 最大似然估计

最大似然估计（maximum likelihood estimation，简记为 MLE）是利用总体的分布密度或概率分布的表达式及其样本所提供的信息求未知参数估计量的一种方法。似然函数并不局限于独立同分布的样本。

- 连续型随机变量：似然函数是样本的联合密度函数
- 离散型随机变量：似然函数是样本的联合概率函数

对于一元线性回归模型参数的最大似然估计，如果已经得到样本观测值 $(x_i, y_i), i = 1, \ldots, n$，那么在假设 $\epsilon_i \sim N(0, \sigma^2)$ 时，y_i 服从如下正态分布：

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

y_i 的分布密度为：

$$f_i(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}[y_i - (\beta_0 + \beta_1 x_i)]^2 \right\}$$

$y_1, y_2, \ldots, y_n$ 的似然函数为：

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f_i(y_i) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\} \end{aligned}$$

取对数似然函数为：

$$\ln(L) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

至此与最小二乘原理相同。

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \text{(有偏估计)}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \text{(无偏估计)}$$

2.3 最小二乘估计的性质

最小二乘估计的性质

- 线性

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} y_i$$

- 无偏性

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

$$\begin{aligned} E(\hat{\beta}_1) &= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} E(y_i) \\ &= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i) = \beta_1 \end{aligned}$$

其中用到 $\sum(x_i - \bar{x}) = 0, \sum(x_i - \bar{x})x_i = \sum(x_i - \bar{x})^2$

- β_0, β_1 的方差

$$var(\hat{\beta}_1) = \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]^2 var(y_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$var(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \sigma^2$$

$$cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{\bar{x}}{L_{xx}} \sigma^2$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \sigma^2 \right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$$

高斯-马尔可夫条件

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

2.3.1 回归方程的显著性检验

t 检验用于检验回归系数的显著性，检验的原假设是：

$$H_0 : \beta_1 = 0$$

对立假设是：

$$H_1 : \beta_1 \neq 0$$

由

$$\hat{\beta}_1 \sim N(\beta_0, \frac{\sigma^2}{L_{xx}})$$

当原假设 $H_0 : \beta_1 = 0$ 成立时，有

$$\hat{\beta}_1 \sim N(0, \frac{\sigma^2}{L_{xx}})$$

构造 t 统计量

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}}$$

其中

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

P − value

- p 值即显著性概率值 Significence Probability Value
- 是当原假设为真时得到目前的样本以及更极端样本的概率，所谓极端就是与原假设相背离
- 它是用此样本拒绝原假设所犯弃真错误的真实概率，被称为观察到的 (或实测的) 显著性水平

利用 p 值进行检验的决策准则

- 若 p 值 $> \alpha$，不能拒绝 H_0
- 若 p 值 $< \alpha$，拒绝 H_0

双侧检验 p 值 = 2× 单侧检验 p 值

F 检验是根据平方和分解式，直接从回归效果检验回归方程的显著性。

平方和分解式是：

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

$$SST=SSR+SSE$$

总的离差平方和：$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$，
回归平方和：$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，
残差平方和：$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$。

构造 F 统计量如下

$$F=\frac{SSR/1}{SSE/(n-2)}$$

相关系数的显著性检验

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\hat{\beta}_1\sqrt{\frac{L_{xx}}{L_{yy}}}$$

r 为 x 与 y 的简单相关系数，简称相关系数。

两变量间相关程度的强弱分为以下几个等级：

- 当 $|r| \geq 0.8$ 时，视为高度相关；
- 当 $0.5 \leq |r| \leq 0.8$，视为低度相关；
- 当 $0.3 \leq |r| \leq 0.5$ 时，视为低度相关；
- 当 $|r| < 0.3$ 时，表明两个变量之间的相关程度极弱，在实际应用中可视为不相关。

对于一元线性回归，这三种检验的结果是完全一致的。

$$H_0 : \beta = 0 \qquad t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}}$$

$$H_0 : \rho = 0 \qquad t = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}}$$

$$H_0 : \text{回归无效} \qquad F = \frac{SSR/1}{SSE/(n-2)}$$

回归平方和与总离差平方和之比定义为决定系数，也称为判定系数、确定系数。记为 R^2

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

可以证明

$$R^2 = \frac{SSR}{SST} = \frac{L_{xy}^2}{L_{xx}L_{yy}} = r^2$$

决定系数 R^2 是一个反映直线与样本观测拟合优度的相对指标，是因变量的变异中能用自变量解释的比例。其数值在 0 ~ 1 之间，可以用百分数表示。

残差：　$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

误差：　$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$

残差项 e_i 是误差项 ϵ 的估计值。

残差的性质

- 性质 1：　$E(e_i) = 0$ 证明：

$$E(e_i) = E(y_i) - E(\hat{y}_i) = (\beta_0 + \beta_1 x_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

- 性质 2：　$var(e_i) = \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}}\right]\sigma^2 = (1 - h_{ii})\sigma^2$

  其中 $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}}$ 称为杠杆值。

- 性质 3：残差满足约束条件

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases}$$

这表明残差 $e_1, e_2, \ldots, e_n$ 的相关的，不是独立的。

2.3.2 回归系数的区间估计

由 $\hat{\beta}_1 \sim N(\beta_0, \frac{\sigma^2}{L_{xx}})$ 可得

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{L_{xx}}}{\hat{\sigma}}$$

服从自由度为 $n-2$ 的 t 分布

$$P(|\frac{(\hat{\beta}_1 - \beta_1)\sqrt{L_{xx}}}{\hat{\sigma}}| < t_{\alpha/2}(n-2)) = 1 - \alpha$$

上式等价于

$$P(\hat{\beta}_1 - t_{\alpha/2}\frac{\hat{\sigma}}{L_{xx}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2}\frac{\hat{\sigma}}{L_{xx}}) = 1 - \alpha$$

即得到 β_1 的置信度为 $1-\alpha$ 的置信区间为：

$$(\hat{\beta}_1 - t_{\alpha/2}\frac{\hat{\sigma}}{L_{xx}}, \hat{\beta}_1 + t_{\alpha/2}\frac{\hat{\sigma}}{L_{xx}})$$

2.3.3 预测和控制

单值预测就是用单个值作为因变量新值的预测值。

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$E(\hat{y}_0) = E(y) = \beta_0 + \beta_1 x_0$$

2.3.4 区间预测

区间预测就是对于给定的显著水平 α，找到一个区间 (T_1, T_2)，使对应于某特定的 x_0 的实际值 y_0 以 $1-\alpha$ 的概率被区间 (T_1, T_2) 包含，用公示表示：

$$P(T_1 < y_0 < T_2) = 1 - \alpha$$

因变量新值的区间预测

首先要给出估计值 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 的分布，在正态性假设下 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 服从正态分布，其期望为 $E(\hat{y}_0) = \beta_0 + \beta_1 x_0$，计算其方差

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 = \sum_{i=1}^{n} [\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}}] y_i$$

$$var(\hat{y}_0) = \sum_{i=1}^{n} [\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}}] var(y_i) = [\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}] \sigma^2$$

从而得

$$\hat{y}_0 \sim N(\beta_0 + \beta_1 x_0, (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}) \sigma^2)$$

记 $h_{00} = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}$

$$\hat{y}_0 \sim N(\beta_0 + \beta_1 x_0, h_{00} \sigma^2)$$

y_0 与 $\hat{y}_0$ 独立

$$var(y_0 - \hat{y}_0) = var(y_0) + var(\hat{y}_0) = \sigma^2 + h_{00} \sigma^2$$

于是

$$y_0 - \hat{y}_0 \sim N(0, (1 + h_{00})\sigma^2)$$

进而可知统计量

$$t = \frac{y_0 - \hat{y}_0}{\sqrt{1 + h_{00}}\hat{\sigma}} \sim t(n-2)$$

$$P(|\frac{y_0 - \hat{y}_0}{\sqrt{1 + h_{00}}\hat{\sigma}}| \leq t_{\alpha/2}(n-2)) = 1 - \alpha$$

由此可以求得 y_0 的置信度为 $1-\alpha$ 的置信区间为：

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{1 + h_{00}}\hat{\sigma}$$

当样本量 n 较大，$|x_0 - \bar{x}|$ 较小时，h_{00} 接近 0，y_0 的置信度为 95% 的置信区间近似为：

$$\hat{y}_0 \pm 2\hat{\sigma}$$

因变量新值的平均值的区间预测

由于$E(y_0) = \beta_0 + \beta_1 x_0$是常数，

$$\hat{y}_0 - E(y_0) \sim N(\beta_0 + \beta_1 x_0, (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}})\sigma^2)$$

可得置信水平为 $1-\alpha$ 的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{h_{00}}\hat{\sigma}$$

思考题

Anscombe(1973)[3]构造了四组数据集，他们具有相同的相关系数。

```
> anscombe <- read.table ('anscombe.txt',header = T)
> head(anscombe)
   Y1 X1    Y2 X2    Y3 X3    Y4 X4
1 8.04 10 9.14 10  7.46 10 6.58  8
2 6.95  8 8.14  8  6.77  8 5.76  8
3 7.58 13 8.74 13 12.74 13 7.71  8
4 8.81  9 8.77  9  7.11  9 8.84  8
```

[3]https://github.com/xuejunma/ar2022/blob/main/anscombe.txt

5 8.33 11 9.26 11 7.81 11 8.47 8
6 9.96 14 8.10 14 8.84 14 7.04 8

- 画出散点图
- 建立线性模型
- 画出拟合图

第 2 讲 练习

1. 小作业
   - p.50 2.8,2.11
   - p.51 2.14 编程计算
2. 大作业

   考虑没有截距项的模型：

   $$y_i = \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

   - β_1 的估计式
   - $SST = SSR + SSE$ 是否成立，说出理由。
   - R^2 一定在大于零吗，说出理由。

第 3 讲 多元线性回归

内容提要

- diabetes 数据
- 最小二乘
- 系数检验
- 模型检验

3.1 diabetes 数据

糖尿病是一种常见的慢性病。diabetes 包含 12 个变量，共有 442 个样。表3.1给出具体的变量名、变量类型和含义说明。我门目的是研究哪些指标对糖尿病有影响。换句话研究 AGE、SEX、BMI、BP 和 S1-S6 对 Y 进行建模分析。这一讲主要利用多元线性模型构建模型。在进行建模前，我们先介绍多元线性模型。

表 3.1: diabetes 数据变量说明

变量名	变量类型	变量含义
AGE	连续	年龄
SEX	分类变量	性别；1 代表男，2 代表女
BMI	连续	身体质量指数，BMI= 体重 (kg) ÷ 身高 (m) 的平方
BP	连续	血压
S1, S2, S3, S4, S5, S6	连续	六种血清的测试量值
Y	连续	测试者一年后糖尿病发展的量化值

表3.2代码

```
> rm(list=ls())
> diabetes<-read.table("diabetes.txt", header=TRUE)
> head(diabetes)
```

3.2 多元线性回归模型的一般形式

设因变量 Y 与自变量 $x_1, x_2, \ldots, x_p$ 的线性回归模型为：

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

表 3.2: diabetes 数据前 6 行

id	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
1	59	2	32.10	101.00	157	93.20	38.00	4.00	4.86	87	151
2	48	1	21.60	87.00	183	103.20	70.00	3.00	3.89	69	75
3	72	2	30.50	93.00	156	93.60	41.00	4.00	4.67	85	141
4	24	1	25.30	84.00	198	131.40	40.00	5.00	4.89	89	206
5	50	1	23.00	101.00	192	125.40	52.00	4.00	4.29	80	135
6	23	1	22.60	89.00	139	64.80	61.00	2.00	4.19	68	97

其中 $\beta_0, \beta_1, \ldots, \beta_p$ 是参数，ε 是误差项。

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

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

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

3.2.1 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

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n) \tag{3.1}$$

笔记 这些基本假设必须要牢记，因为后面内容，我们将讨论违背这些假设的处理方法。当

然，这假设也可以释放，但需要更高级的模型。

在正态假定下，我们重写模型的表达:

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n) \tag{3.2}$$

这个表达式可能会引起误解。我们必须明白这个表达式的前提是误差项服从正态分布。后面在广义线性模型中，我们将再继续讨论。从上面的表达式，我们可以得到。

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

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

3.3 回归参数的估计

这一部分讨论参数 $\boldsymbol{\beta}$ 的估计方法，常见的有最小二乘法和极大似然的方法。相比后者，前者对误差项没有分布要求，并且不需要 ε_i 相互独立。换句话说最小二乘法需要基本假设（1）成立，再加上 $E\varepsilon_i = 0$; 极大似然估计需要基本假设全部成立。

3.3.1 最小二乘估计

最小二乘估计使得下面目标函数最小，

$$Q(\beta_0,\beta_1,\cdots,\beta_p)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2 \tag{3.3}$$

目标函数关于参数求到，得到下面表达式：

$$\begin{cases}\dfrac{\partial Q}{\partial\beta_0}=-2\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})\\ \dfrac{\partial Q}{\partial\beta_1}=-2\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_p)x_{i1}\\ \dfrac{\partial Q}{\partial\beta_2}=-2\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_p)x_{i2}\\ \quad\vdots\\ \dfrac{\partial Q}{\partial\beta_p}=-2\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_p)x_{ip}\end{cases}$$

为了方便表达，我们采用矩阵的形式，即

$$Q(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

关于 $\boldsymbol{\beta}$ 求导

$$\boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}$$

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

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \tag{3.4}$$

此时，最小二乘的估计我们变得到了。我们可以使用 R 的函数 solve 函数求矩阵的拟。需要说明的是矩阵的求拟不是一件容易的事情。很早以前 solve 的方法利用线性方程组求解，但这种方法很慢当自变量个数很大，后面修改为矩阵分解的方法得到。这一部分的讨论相见附录3.11。

下面，我们介绍 σ^2 的估计。观测值 y_i 得回归拟合值

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

矩阵的表达式为

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

其中 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ 为帽子矩阵, 其主对角线元素记为 h_{ii}. 这个矩阵将在下面内容反复出现，请大家记住它的样子。下面介绍它的性质:

(1) $\boldsymbol{H}^\top = \boldsymbol{H}$

(2) $\boldsymbol{H}^2 = \boldsymbol{H}$

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

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

这说明 $\boldsymbol{H}$ 是对称幂等矩阵。想想幂等矩阵的性质。[1]

残差是观测值减去拟合值，我们通常如下定义:

$$\boldsymbol{e} = (e_1, e_2, \cdots, e_n)^\top = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

[1]如果 $\boldsymbol{A}$ 幂等矩阵，则 (1) rank(A)=tr(A) （2）I-A 也是幂等矩阵。这里仅仅列举了两个，后面主要利用这两个性质; 其他的性质大家可以查查资料。

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

进而得到

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

又因为

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

从而，我们得到 σ^2 的估计量为:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-p-1}(\boldsymbol{e}^{\top}\boldsymbol{e}) = \frac{RSS}{n-p-1}$$

其中 RSS 是残差平方和 (Residual Sum of Squares)。

3.3.2 极大似然估计

极大似然基于分布得到，从（3.4），我们可以得到 $\boldsymbol{Y}$ 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

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

两边同时取对数似然

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

等价于使 $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ 达到最小，这又与最小二乘法相同。

虽然 $\boldsymbol{\beta}$ 的极大似然估计与最小二乘估计一样，但是极大似然得到的

$$\sigma^2_{ML} = \frac{RSS}{n}.$$

相比最小二乘估计，σ^2_{ML} 是有偏估计。所以，σ^2 常常采用最小二乘估计。这里，我们需要注意有偏估计不是不好，只是在无偏评价的标准下不好，不代表其他评价标准下不好，比如均方误差标准下，有偏估计可能优于无偏估计。

3.3.3 参数估计量的性质

参数的估计性质的讨论需要上面的三个基本假设条件。不同的性质需要假设的条件不同。

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

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

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

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

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

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

- 性质 4：Gauss-Markov 定理
  预测函数 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20} + \cdots + \hat{\beta}_p x_{p0}$ 是 $\hat{\boldsymbol{\beta}}$ 的线性函数。
  希望 $\hat{\boldsymbol{\beta}}$ 的线性函数波动越小越好！

  下面我们考虑 $\boldsymbol{\beta}$ 一类估计，其是 Y 的线性组合：

$$\widetilde{\beta} = \boldsymbol{A}^\top\boldsymbol{Y}$$

  其中 $\boldsymbol{A}$ 是 $\boldsymbol{X}$ 的函数。最小二乘估计是其特例（$\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. 存不存在最佳的 $\boldsymbol{A}$？Gauss-Markov 定理讲给出我们答案。

Gauss-Markov 定理：

在假定 $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}, D(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$ 时，$\boldsymbol{\beta}$ 的任一线性函数 $\boldsymbol{c}^\top\boldsymbol{\beta}$ 的最小方差线性无偏估计（BLUE）为 $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$，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 $\boldsymbol{c}^\top\boldsymbol{\beta}$ 的无偏估计，比最小二乘估计 $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$ 的方差更小。

(3) 可能存在 $\boldsymbol{c}^\top\boldsymbol{\beta}$ 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 $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$ 更好。

(4) 在正态假定下，$\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{c}^\top\boldsymbol{\beta}$ 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 $\boldsymbol{c}^\top\boldsymbol{\beta}$ 的无偏估计，比最小二乘估计 $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$ 方差更小。

- 在线性回归中，在给定 $p+1$ 维的常向量 $\boldsymbol{c}$，则在 $\boldsymbol{c}^\top\boldsymbol{\beta}$ 的所有线性无偏估计中，最小二乘 $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$ 唯一具有最小方差估计。[2]

[2]梅长林，王宁。近代回归分析方法，2012，中国科学出版社.p.7.

证明：设 $\boldsymbol{\alpha}^\top \boldsymbol{Y}$ 为 $\boldsymbol{c}^\top \boldsymbol{\beta}$ 的任意一个线性无偏估计，于是对于一切的 $\boldsymbol{\beta}$，有

$$\boldsymbol{c}^\top \boldsymbol{\beta} = \mathbf{E}(\boldsymbol{\alpha}^\top \boldsymbol{Y}) = \boldsymbol{\alpha}^\top \boldsymbol{X}\boldsymbol{\beta}$$

由于 $\boldsymbol{\beta}$ 的任意性可得 $\boldsymbol{c}^\top = \boldsymbol{\alpha}^\top \boldsymbol{X}$。任意一个向量 $||\boldsymbol{Z}||^2 = \boldsymbol{Z}^\top \boldsymbol{Z}$，则 $||\boldsymbol{Z}||^2 = 0$ 当且仅当 $\boldsymbol{Z} = 0$。

$$\begin{aligned} Var(\boldsymbol{\alpha}^\top \boldsymbol{Y}) &= \boldsymbol{\alpha}^\top Var(\boldsymbol{Y})\boldsymbol{\alpha} = \sigma^2 ||\boldsymbol{\alpha}||^2 \\ &= \sigma^2 ||\boldsymbol{\alpha} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{c} + \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{c}||^2 \\ &= \sigma^2 ||\boldsymbol{\alpha} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{c}||^2 + \sigma^2 \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{c} \\ &+ 2\sigma \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top (\boldsymbol{\alpha} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{c}) \end{aligned}$$

将 $\boldsymbol{c}^\top = \boldsymbol{\alpha}^\top \boldsymbol{X}$ 代入第三项得此项为零。$Var(\boldsymbol{c}^\top \hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{c}$，故 $Var(\boldsymbol{\alpha}^\top \boldsymbol{Y}) \geq Var(\boldsymbol{c}^\top \hat{\boldsymbol{\beta}})$ 等号当且仅当 $\boldsymbol{\alpha} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{c} = 0$ 成立，即 $\boldsymbol{\alpha}^\top \boldsymbol{Y} = \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{c}^\top \hat{\boldsymbol{\beta}}$

- 性质 5 $cov(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = 0$

  此性质说明 $\hat{\boldsymbol{\beta}}$ 与 $\boldsymbol{e}$ 不相关，在正态假定下 $\hat{\boldsymbol{\beta}}$ 与 $\boldsymbol{e}$ 等价于 $\hat{\boldsymbol{\beta}}$ 与 $\boldsymbol{e}$ 独立，从而 $\hat{\boldsymbol{\beta}}$ 与 $SSE = \boldsymbol{e}^\top \boldsymbol{e}$ 独立。

- 性质 6：当 $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$ （等价$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$）时，则

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

(2) $SSE/\sigma^2 \sim \chi^2(n-p-1) \Rightarrow E(SSE) = \sigma^2(n-p-1)$

(3) $\hat{\boldsymbol{\beta}}$ 与 SSE/σ^2 相互独立。

证明 (2)

由于 $e = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})$=$(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}$，从而 $SSE = \boldsymbol{\varepsilon}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}$ 再由于 $\boldsymbol{I}-\boldsymbol{H}$ 是对称幂等矩阵，故 $rank(\boldsymbol{I}-\boldsymbol{H}) = n-p-1$，从而存在 n 阶正交矩阵 $\boldsymbol{P}$，使得

$$\boldsymbol{P}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{P} = \begin{pmatrix} \boldsymbol{I}_{n-p-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}$$

令

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top = \boldsymbol{P}^\top\boldsymbol{\varepsilon} \Rightarrow \boldsymbol{\varepsilon} = \boldsymbol{P}\boldsymbol{\eta}$$

则 $E(\boldsymbol{\eta}) = 0, Var(\boldsymbol{\eta}) = \sigma^2\boldsymbol{P}^\top\boldsymbol{P} = \sigma^2\boldsymbol{I}$，从而 $\boldsymbol{\eta} \sim N(0, \sigma^2\boldsymbol{I})$，进而

$$\frac{1}{\sigma^2}SSE = \frac{1}{\sigma^2}\boldsymbol{\eta}^\top\boldsymbol{P}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{P}\boldsymbol{\eta} = \frac{1}{\sigma^2}(\eta_1^2 + \ldots + \eta_{n-p-1}^2) \tag{3.5}$$

由于 $\eta_i \sim N(0, \sigma^2)$ 相互独立，并且 $\eta_i/\sigma \sim N(0,1)$，从而 $\eta_i^2/\sigma^2 \sim \chi^2(1)$，再由卡方分布得可加性得 $SSE/\sigma^2 \sim \chi^2(n-p-1)$

证明 (3)

将 n 阶正交矩阵 $\boldsymbol{P}$ 分成两块 $\boldsymbol{P} = (\boldsymbol{P}_1, \boldsymbol{P}_2)$，其中 $\boldsymbol{P}_1$ 由 $\boldsymbol{P}$ 的前 $n-p-1$ 列组成，$\boldsymbol{P}_2$ 后

$p+1$ 列组成，从而

$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{P}_1^\top \\ \boldsymbol{P}_2^\top \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{P}_1^\top \boldsymbol{\varepsilon} \\ \boldsymbol{P}_2^\top \boldsymbol{\varepsilon} \end{pmatrix}$$

由于 $\boldsymbol{\eta} \sim N(0, \sigma^2 \boldsymbol{I})$，所以 $\boldsymbol{\eta}_1 \sim N(0, \sigma^2 \boldsymbol{I}_{n-p-1})$，$\boldsymbol{\eta}_2 \sim N(0, \sigma^2 \boldsymbol{I}_{p+1})$，且 $\boldsymbol{\eta}_1$ 和 $\boldsymbol{\eta}_2$ 相互独立

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P} \boldsymbol{\eta} \triangleq \boldsymbol{D} \boldsymbol{\eta}$$

其中 $\boldsymbol{D} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P} = (\boldsymbol{D}_1, \boldsymbol{D}_2)$, $\boldsymbol{D}_1$ 由 $\boldsymbol{D}$ 的前 $n-p-1$ 列组成。

$$(\boldsymbol{D}_1, \boldsymbol{D}_2) \begin{pmatrix} \boldsymbol{I}_{n-p-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} = \boldsymbol{D} \boldsymbol{P}^\top (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{P} = \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{P} = 0$$

从而 $\boldsymbol{D}_1 = 0$，所以

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{D}_1, \boldsymbol{D}_2) \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} = \boldsymbol{D}_2 \boldsymbol{\eta}_2$$

从等式 (3.5)，可以得

$$SSE = \boldsymbol{\eta}_1^\top \boldsymbol{\eta}_1$$

由于 $\boldsymbol{\eta}_1$ 和 $\boldsymbol{\eta}_2$ 相互独立，从而 SSE 和 $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ 相互独立。

- 性质 7：当 $\boldsymbol{Y}\sim N(\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n)$（等价$\boldsymbol{\varepsilon}\sim N(0,\sigma^2\boldsymbol{I})$）时，则

  (1) $E(\boldsymbol{e})=0$

  (2) $Var(\boldsymbol{e})=\sigma^2(\boldsymbol{I}-\boldsymbol{H})$

  (3) $\boldsymbol{e}\sim N(\boldsymbol{0},\sigma^2(\boldsymbol{I}-\boldsymbol{H}))$

3.4 约束的最小二乘

这一部分，我们主要讨论带约束的最小二乘。这个方法以应比较广泛，特别是成分数据分析（这一部分见附录）。

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$$

约束条件

$$\boldsymbol{A\beta}=\boldsymbol{b}$$

其中 $\boldsymbol{A}$ 是 $m\times(p+1)$ 阶行满秩常值矩阵。$\boldsymbol{\varepsilon}\sim N(0,\sigma^2\boldsymbol{I})$，这个条件可弱一些，最小二乘法的条件即可。

- 如何估计?

利用 Lagrance 方法求解：设 $\boldsymbol{\lambda}$ 为 Lagrance 乘子，构造辅助函数

$$\begin{aligned} F(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\top}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}^{\top}(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{b}) \\ &= \boldsymbol{Y}^{\top}\boldsymbol{Y} - 2\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{Y} + \boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}^{\top}\boldsymbol{A}^{\top}\boldsymbol{\lambda} - 2\boldsymbol{b}^{\top}\boldsymbol{\lambda} \end{aligned}$$

求偏导

$$\frac{\partial F(\boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^{\top}\boldsymbol{Y} + 2\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{A}^{\top}\boldsymbol{\lambda} = 0$$

$$\frac{\partial F(\boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 2(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{b}) = 0$$

进而得到

$$\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}^{\top}\boldsymbol{\lambda} = \boldsymbol{X}^{\top}\boldsymbol{Y} \tag{3.6}$$

$$\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b} \tag{3.7}$$

令 $\hat{\boldsymbol{\beta}}_c$ 和 $\hat{\boldsymbol{\lambda}}_c$ 为上述方程组的解，

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y} - (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{A}^{\top}\hat{\boldsymbol{\lambda}}_c$$

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{A}^\top \hat{\boldsymbol{\lambda}}_c$$

将代入 (3.6) 的第二等式 $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$，从而

$$\hat{\boldsymbol{\lambda}}_c = [\boldsymbol{A}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{A}^\top]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})$$

可以得到

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{A}^\top [\boldsymbol{A}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{A}^\top]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})$$

为了确保 $\hat{\boldsymbol{\beta}}_c$ 的确为约束条件下的 $\boldsymbol{\beta}$ 的最小二乘估计，我们还需证明下面两点（**作业**）

- (1)$\boldsymbol{A}\hat{\boldsymbol{\beta}}_c = \boldsymbol{b}$
- (2) 对一切满足条件的 $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$ 的 $\boldsymbol{\beta}$，都有 $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \geq ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_c||^2$

3.5 模型评价

常见的模型评价指标主要有三种：模型显著性检验、系数显著性检验和决定系数。模型显著性检验主要检验整个模型的是否显著，如果不显著，说明模型不合适。系数显著性检验主要针对每一个系数和常数项而言。只要有一个系数显著，前面的模型检验便显著。决定系

数越接近 1，说明模型拟合的越好。

另外，残差分析也是模型拟合检验的重要方式。模型显著性检验、系数显著性检验前提是线性回归的假设成立。而残差分析是检验假设是否成立。这一点大家需要多注意。残差分析将在下一讲详细阐述。

3.6 残差分析

我们首先介绍残差，因为残差在模型评价中非常重要。

我们首先给出一种回归 LOO 回归（Leave-one-out Regression) 是去掉某一个观察值 i 剩下样本的参数估计值

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\sum_{j\neq i} X_j X_j^{\top}\right)^{-1}\left(\sum_{j\neq i} X_j Y_j\right) \\ &= (\boldsymbol{X}^{\top}\boldsymbol{X} - X_i X_i^{\top})^{-1}(\boldsymbol{X}^{\top}\boldsymbol{Y} - X_i Y_i) \\ &= \left(\boldsymbol{X}_{(-i)}^{\top}\boldsymbol{X}_{(-i)}\right)^{-1}\left(\boldsymbol{X}_{(-i)}^{\top}\boldsymbol{Y}_{(-i)}\right)\end{aligned}$$

其中 $\boldsymbol{X}_{-i}$ 和 $\boldsymbol{Y}_{-i}$ 分别表示去掉第 i 个观测值的数据矩阵。

如果记 $\widetilde{\varepsilon}_i = Y_i - \widetilde{Y}_i$，其中 $\widetilde{Y}_i = \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_{(-i)}$，

定理 3.1

1.

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X_i \widetilde{\varepsilon}_i$$

2.

$$\widetilde{\varepsilon}_i = (1 - h_{ii})^{-1} e_i$$

♡

这个证明留在联系。

对于 σ^2 的估计，一个自然的估计量是

$$\widetilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

由于

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \\ &= (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{Y} \\ &= \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{M}\boldsymbol{\varepsilon}\end{aligned}$$

其中 $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{H}$.

从而

$$\widetilde{\sigma}^2 = \frac{1}{n}\boldsymbol{e}^\top\boldsymbol{e} = \frac{1}{n}\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}$$

进而

$$
\begin{aligned}
E(\hat{\sigma}^2|\boldsymbol{X}) &= E\left(\frac{1}{n}\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon}\right) \\
&= E\left[tr\left(\frac{1}{n}\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon}\right)|\boldsymbol{X}\right] \\
&= E\left[tr\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right)|\boldsymbol{X}\right] \\
&= \frac{1}{n}tr\left[E(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)|\boldsymbol{X}\right] \\
&= \frac{1}{n}tr(\boldsymbol{M}\sigma^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})\sigma^2 \\
&= \frac{n-p-1}{n}\sigma^2
\end{aligned}
$$

这说明 $\widetilde{\sigma}^2$ 是有偏估计，由此引入无偏估计

$$
\widehat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}e_i^2
$$

另外，还有一种估计是基于标准化的残差：

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{\sqrt{1-h_{ii}}}\right)^2$$

3.6.1 模型显著性检验

模型的显著性检验针对整个模型，不是具体的系数，其原假设为：

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{3.8}$$

如果原假设成立，模型简化为

$$y_i = \beta_0 + \varepsilon_i.$$

此时，我们得到

$$\hat{\beta}_0 = \arg\min \sum_{i]=1}^{n} (y_i - \beta_0)^2$$

经过简单计算，我们得到 $\hat{\beta}_0 = \bar{y}$, 进而我们可以到目标函数为

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

它便是我们经常说的总离差平方和 (SST, Sum of squares Tolal)。一般情况下，SST 可以分解说

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.9}$$

其中 $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ 称为回归平方和 (SST, Sum of squares regression)，$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 称为残差平方和 (SSE, Sum of squares residual)。之所以是回归平方和由于 $\hat{y}_i$ 实际上是条件均值的估计。(3.9) 简单写为：

$$SST = SSR + SSE$$

笔记 需要注意的是 (3.9) 并不是总是成立。如果模型不含有常数项，(3.9) 可能不成立。具体见练习题。

针对 (3.8)，通常构造 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，具体如表3.3。

表 3.3: 模型检验

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			

下面我们介绍约束最小二乘估计的检验。

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0, \sigma \boldsymbol{I})$$

假设

$$H_0 : \boldsymbol{A\beta} = \boldsymbol{b}$$

其中 $\boldsymbol{A}$ 是 $m \times (p+1)$ 阶行满秩常值矩阵。令

$$SSE = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

$$SSE(H_0) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(H_0))^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(H_0))$$

定理 3.2

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

- $SSE/\sigma^2 \sim \chi^2_{n-p-1}$
- 若 H_0 成立，则 $(SSE(H_0) - SSE)/\sigma^2 \sim \chi^2_m$
- SSE 和 $SSE(H_0) - SSE$ 相互独立
- 若 H_0 成立

$$F = \frac{(SSE(H_0) - SSE)/m}{SSE/(n-p-1)} \sim F(m, n-p-1)$$

♡

笔记

- 我们知道 SST 实际是只含有常数项的回归模型的 SSR，还句话说，(3.9) 中的 SSR 等于只含有常数项的回归模型的 SSR 减去 (3.9) 中的 SSE。表3.3 中的 F 检验是上面定理的一个特例，只要假设 $\boldsymbol{A} = (0, 1, 1, \ldots, 1)^\top$ 和 $\boldsymbol{b} = 0$。
- 这个定理的证明类似于性质 6。留作思考题。

3.6.2 系数检验

原假设

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

已知

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

记

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

由此可以构造 t 统计量

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

其中

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

回归方程的显著性检验: 偏 F 统计量

从另外一个角度考虑自变量 x_j 的显著性。

- y 对自变量 $x_1, x_2, \ldots, x_p$ 线性回归的残差平方和为 SSE，回归平方和为 SSR
- 在剔除掉 x_j 后，用 y 对其余的 $p-1$ 个自变量做回归，记所得的残差平方和为 $SSE_{(j)}$，回归平方和为 $SSR_{(j)}$，则自变量 x_j 对回归的贡献为 $\Delta SSR_{(j)} = SSR - SSR_{(j)}$，称为 x_j 的偏回归平方和。由此构造偏 F 统计量

$$F_j = \frac{\Delta SSR_{(j)}/1}{SSE/(n-p-1)}$$

- 当原假设 $H_{0j}: \beta_j = 0$ 成立时，偏 F 统计量 F_j 服从自由度为 $(1, n-p-1)$ 的 F 分布，此 F 检验 t 检验是一致的，可以证明 $F_j = t_j^2$。

回归系数的置信区间

因为

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t(n-p-1)$$

可得 β_j 的置信度为 $1-\alpha$ 的置信区间为

$$(\hat{\beta}_j - t_{\alpha/2}\sqrt{c_{jj}}\hat{\sigma}, \hat{\beta}_j + t_{\alpha/2}\sqrt{c_{jj}}\hat{\sigma})$$

3.6.2.1 拟合优度

- 定义样本的决定系数

$$R^2 = 1 - \frac{SSE}{SST}$$

- y 关于 $x_1, x_2, \ldots, x_p$ 的样本复相关系数

$$R = \sqrt{R^2} = \sqrt{\frac{SSR}{SST}}$$

笔记

- 样本决定系数 R^2 越大，说明回归方程拟合原始数据 y 的观测值的效果越好。
- 但由于 R^2 的大小与样本容量 n 以及自变量个数 p 有关，当 n 与 p 的数目接近时，R^2 容易接近于 1，这说明 R^2 中隐含着一些虚假成分
- 仅由 R^2 的值很大，去推断模型优劣一定要慎重。

- 调整 R^2

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

通常用来比较自变量个数不同的模型拟合效果，不能理解为变量 Y 的总方差中能由自变量解释的比例。

下面，我们从另一个角度尝试解释决定系数。

$$\begin{aligned} R^2 &= 1 - \frac{SSE}{SST} \\ &= 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \\ &= 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_Y} \end{aligned}$$

从而，R^2 是 ρ^2 的一个估计

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} = \frac{Var(\boldsymbol{X}_i^\top \boldsymbol{\beta})}{Var(Y_i)}$$

我们知道 $\hat{\sigma}^2$ 和 $\hat{\sigma}_Y$ 均是有偏估计，因此，我们进行无偏估计代替，从而产生了调整的 R^2

（Adjusted R-squared）

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

3.7 中心化和标准化

多元线性回归模型的经验回归方程为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

中心化回归系数

经过样本中心 $(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p; \bar{y})$，将坐标原点移到样本中心，即

$$x'_{ij} = x_{ij} - \bar{x}_j, \quad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, p$$

$$y'_i = y_i - \bar{y}, \quad i = 1, 2, \ldots, n$$

上述经验方程转换为

$$\hat{y}' = \hat{\beta}_1 x'_1 + \hat{\beta}_2 x'_2 + \cdots + \hat{\beta}_p x'_p$$

回归常数项为

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 - \cdots - \hat{\beta}_p\bar{x}_p$$

标准化回归系数

样本数据的标准化公式为

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{L_{jj}}}, \quad i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, p$$

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{L_{yy}}}, \quad i = 1, 2, \cdots, n$$

其中

$$L_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

标准化样本的经验回归方程为

$$\hat{y}^* = \hat{\beta}_1^* x_1^* + \hat{\beta}_2^* x_2^* + \cdots + \hat{\beta}_p^* x_p^*$$

$$\hat{\beta}_j^* = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j, \quad j = 1, 2, \cdots, p$$

3.8 相关阵与偏相关系数

样本相关阵

r_{ij} 为 x_i 与 x_j 之间简单的相关系数，自变量样本相关阵

$$\boldsymbol{r} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

同时，相关阵可表示为：

$$\boldsymbol{r} = (\boldsymbol{X}^*)'\boldsymbol{X}^*$$

$\boldsymbol{X}^* = (x_{ij}^*)_{n\times p}$ 表示中心标准化的设计阵。

增广的样本相关阵为：

$$\boldsymbol{r} = \begin{pmatrix} 1 & r_{y1} & r_{y2} & \cdots & r_{yp} \\ r_{1y} & 1 & r_{12} & \cdots & r_{1p} \\ r_{2y} & r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{py} & r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

偏相关系数和偏决定系数

- 当其他变量被固定后，给定的任两个变量之间的相关系数，叫**偏相关系数**。
  偏相关系数可以度量 $p+1$ 个变量 $y, x_1, x_2, \cdots x_p$ 之中任意两个变量的线性相关程度，而这种相关程度是在固定其余 $p-1$ 个变量的影响下的线性相关
- 偏判定系数测量在回归方程中已包含若干个自变量时，再引入某一个新的自变量后 y 的剩余变差的相对减少量，它衡量 y 的变差减少的边际贡献。

两个自变量的偏判定系数

二元线性回归模型为：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, 2, \cdots, n$$

记 $SSE(x_2)$ 是模型中只含有自变量 x_2 时 y 的残差平方和，$SSE(x_1 x_2)$ 是模型中同时含有自变量 x_1 和 x_2 时 y 的残差平方和。因此模型中已含有 x_2 时再加入 x_1 使 y 的剩余变差的相对减小量为

$$R^2_{y1;2} = \frac{SSE(x_2) - SSE(x_1, x_2)}{SSE(x_2)}$$

此即模型中已含有 x_2 时，y 与 x_1 的偏判定系数。

同样，模型中已含有 x_1 时，y 与 x_2 的偏判定系数为：

$$R^2_{y2;2} = \frac{SSE(x_1) - SSE(x_1, x_2)}{SSE(x_1)}$$

$$r^2_{y1;2,\cdots,p} = \frac{SSE(x_2, \cdots, x_p) - SSE(x_1, x_2, \cdots, x_p)}{SSE(x_2, \cdots, x_p)}$$

偏决定系数与回归系数显著性检验的偏 F 值是等价的。

偏判定系数的平方根称为偏相关系数，其符号与相应的回归系数的符号相同。偏相关系数与回归系数显著性检验的 t 值是等价的。

固定 $x_3, \cdots, x_p$ 保持不变时，x_1 与 x_2 之间的偏相关系数为：

$$r_{12;3,\cdots,p} = \frac{-\Delta_{12}}{\sqrt{\Delta_{11} \cdot \Delta_{22}}}$$

其中符号 Δ_{ij} 表示相关矩阵 $(r_{ij})_{p\times p}$ 第 i 行第 j 列元素的代数余子式。验证以下关系

$$r_{12;3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

3.9 diabetes 数据分析

diabetes 分析的代码

```
rm(list=ls())
library(xtable)
diabetes<-read.table("/Users/yinuo/Desktop/regressionbook/diabetes.txt", header=TRUE)
head(diabetes, 20)
xtable::xtable(head(diabetes,
```

```
                20),caption = "diabetes", label = 'diabetestable1')

fit_lm <- lm(Y~., data=diabetes)
summary(fit_lm)

Call:
lm(formula = Y ~ ., data = diabetes)

Residuals:
     Min       1Q   Median       3Q      Max
-155.827  -38.536   -0.228   37.806  151.353

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -334.56714   67.45462  -4.960 1.02e-06 ***
AGE           -0.03636    0.21704  -0.168 0.867031
```

```
SEX          -22.85965    5.83582  -3.917 0.000104 ***
BMI            5.60296    0.71711   7.813 4.30e-14 ***
BP             1.11681    0.22524   4.958 1.02e-06 ***
S1            -1.09000    0.57333  -1.901 0.057948 .
S2             0.74645    0.53083   1.406 0.160390
S3             0.37200    0.78246   0.475 0.634723
S4             6.53383    5.95864   1.097 0.273459
S5            68.48312   15.66972   4.370 1.56e-05 ***
S6             0.28012    0.27331   1.025 0.305990
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.15 on 431 degrees of freedom
Multiple R-squared:  0.5177,    Adjusted R-squared:  0.5066
F-statistic: 46.27 on 10 and 431 DF,  p-value: < 2.2e-16
```

结果显示

3.10 小结和评注

极大似然估计方法在线性回归中似乎没有，其实其有用。后面我们讲接广义线性模型时，估计方法主要是极大似然法。马等（2014）提出一种的回归-核密度回归（Kernel density regression），其使用核密度估计密度函数，然后基于极大似然估计的方法。换句话说密度函数不知，我们可以将其估计出来，进而得到似然函数。具体的见文献。这种方法的缺点是求解比较复杂，需要优化非线性目标函数。至于核密度估计，我们将在后面讲解。

3.11 附录 A

矩阵 QR 分解

设 $\boldsymbol{A} \in \mathbb{R}^{n\times m}$，则存在 $\boldsymbol{Q} \in \mathbb{R}^{n\times p}$，且 $\boldsymbol{Q}^{\top}\boldsymbol{Q} = \boldsymbol{I}_p$，以及一个上三角矩阵 $\boldsymbol{R} \in \mathbb{R}^{p\times p}$ 使得

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$$

奇异值分解代码

```
> rm(list=ls())
> A <- matrix(1:6, 2, 3)
> A.qr <- qr(A)
> Q <- qr.Q(A.qr)
> crossprod(Q)
             [,1]         [,2]
[1,] 1.000000e+00 5.551115e-17
[2,] 5.551115e-17 1.000000e+00
> R <- qr.R(A.qr)
```

```
> Q %*% R
  [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

线性回归求解是一个技术问题，参数估计表达式 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y}$ 逆的解法比较复杂，假设

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$$

则

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y} \\
&= (\boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{R})^{-1}\boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Y} \\
&= (\boldsymbol{R}^\top \boldsymbol{R})^{-1}\boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Y} \\
&= \boldsymbol{R}^{-1}\boldsymbol{R}^{-1\top}\boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Y} \\
&= \boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{Y}
\end{aligned}$$

第三步使用逆的性质 $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$。这样大大降低了计算成本。上三角矩阵的逆求解比

普通矩阵的逆好很多。下面我们模拟数据看看这个计算效果。我们比较了 lm 和 QR 分解，发现结果一致。原因是前者的求解是基于 QR 分解得到的（是否？查看？rm 文档）

QR 代码

```
> rm(list=ls())
> library(MASS)
> ###data begin
> p <- 5
> n <- 100
> mu <- rep(0, p)
> sigm <- diag(p)
> x <- MASS::mvrnorm(n=n, mu=mu, Sigma=sigm)
> beta <- 1 + runif(p, min = 0, max = 1)
> y <- x %*% beta + rnorm(n)
> ###data end
>
> ###X QR begin
```

```
> x.qr <- qr(x)
> Q <- qr.Q(x.qr)
> R <- qr.R(x.qr)
> ###X QR end
>
> beta_hat <- solve(R) %*% t(Q) %*% y
> beta_ols <- coef(lm(y~-1))
> cbind(beta_ols, beta_hat)
```

3.12 附录 B

成分数据是一类性质比较复杂的数据，传统上被定义为约束型数据，通常以比例或百分比来表示。实际上，我们可以直观的认为成分数据是带有相对信息的观测值，其中比例仅仅代表一种可能的表示。同样地，成分数据的相关信息包含在成分之间的比率中。

设 $Y = (Y_1, \ldots, Y_n)^\top$ 是 n 维列向量，成分协变量矩阵 X 是 $n \times p$ 维的，X 的每一行所在的空间是单形，即 $S^p = \left\{(X_1, X_2, ..., X_p) | x_i > 0, i = 1, 2, ..., p; \sum_{i=1}^{p} X_i = 1\right\}$. 由于定和约束

的存在，成分数据各分量之间存在完全多重共线性。Aitchison 和 Bacon (1984) 提出将对数比变换应用于成分协变量，得到线性对数对比模型：

$$Y = Z^p \beta_{\backslash p} + \epsilon, \tag{3.10}$$

其中，$Z^p = \left\{\log(X_{ij}/X_{ip})\right\}$ 是 $n \times (p-1)$ 的矩阵，第 p 个成分是参考元。$\beta_{\backslash p} = (\beta_1, \ldots, \beta_{p-1})^\top$，$\epsilon$ 是 n 维独立同分布的误差项。显然，$p = 1$ 与 $p = 2$ 时得到的 β 是不同的，我们还面临一个 β 选择的问题。参照元如果选的不合适，会影响统计推断的结果。为了解决这个问题，Lin 等 (2014) 引入约束 $\beta_p = -\sum_{j=1}^{p-1} \beta_j$，将模型 (3.10) 转化为如下对称的形式

$$Y = Z\beta + \epsilon, \quad \sum_{j=1}^{p} \beta_j = 0, \tag{3.11}$$

其中，$Z = (Z_{ij}) = \{\log X_{ij}\}_{n \times p}$，$\beta = (\beta_1, ..., \beta_p)^\top$。在公式 (3.11) 基础上，Lin 等 (2014) 建立成分数据的回归模型如下

$$\begin{aligned} \hat{\beta} =& \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - Z_i^\top \beta)^2, \\ & \text{s.t.} \sum_{j=1}^{p} \beta_j = 0, \end{aligned} \tag{3.12}$$

其中，s.t. 代表约束条件 (subject to). 这是带有约束最小二乘的一个特例。

3.13 参考文献

- Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. Biometrika, 71(2): 323-330.
- Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. Biometrika, 101(4): 785-797.
- Chen, X, Ma, X. Zhou, W. (2020). Kernel density regression. Journal of Statistical Planning and Inference. 205, 318-329.

第 3 讲 练习

1. 考虑没有截距项的模型：

$$y_i = \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

   (a). β_1 的估计式

   (b). *SST*=*SSR*+*SSE* 是否成立，说出理由

(c). R^2 一定大于零吗，说明理由。另外，尝试使用模拟实验的说明。

2. Anscombe(1973) 构造了四组数据集 anscombe，请完成下面。
   (a). 画出散点图
   (b). 建立线性模型，并且进行模型检验
   (c). 画出拟合图
   (d). 谈谈您的收获
3. 如果样本矩阵分成两块 $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$，对应的参数也分为两块 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$，则回归模型可以写为

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

假设

$$Q(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2)^\top(\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2)$$

根据前面内容，我们可以得到

$$(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \underset{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}{\texttt{argmin}}\ Q(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \tag{3.13}$$

参数也可以通过下面的方法求解

$$\begin{cases} \boldsymbol{\beta}_1 = \underset{\boldsymbol{\beta}_1}{\texttt{argmin}}\left(\min\limits_{\boldsymbol{\beta}_2} Q(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\right) \\ \boldsymbol{\beta}_2 = \underset{\boldsymbol{\beta}_2}{\texttt{argmin}}\left(\min\limits_{\boldsymbol{\beta}_1} Q(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\right) \end{cases} \tag{3.14}$$

证明：

(a). (3.14) 解的表达式为：

$$\boldsymbol{\beta}_1 = (\boldsymbol{X}_1^\top \boldsymbol{M}_2 \boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1^\top \boldsymbol{M}_2 \boldsymbol{Y})$$
$$\boldsymbol{\beta}_2 = (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{Y})$$

其中 $\boldsymbol{M}_k = \boldsymbol{I}_n - \boldsymbol{X}_k(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top$

(b). 上面的求解与 (3.13) 求解等价。[3]

(c). 看到上面的结论，您有何感想？

[3]分块矩阵逆的计算 $\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BC^{-1}D)^{-1} & -(A - BC^{-1}D)^{-1}BD^{-1} \\ -(D - BA^{-1}C)^{-1}CA^{-1} & (D - BA^{-1}C)^{-1} \end{pmatrix}$

4. LOO 回归（Leave-one-out Regression) 是去掉某一个观察值 i 剩下样本的参数估计值

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\sum_{j\neq i} X_j X_j^\top\right)^{-1}\left(\sum_{j\neq i} X_j Y_j\right)\\ &= (\boldsymbol{X}^\top\boldsymbol{X} - X_i X_i^\top)^{-1}(\boldsymbol{X}^\top\boldsymbol{Y} - X_i Y_i)\\ &= \left(\boldsymbol{X}_{(-i)}^\top\boldsymbol{X}_{(-i)}\right)^{-1}\left(\boldsymbol{X}_{(-i)}^\top\boldsymbol{Y}_{(-i)}\right)\end{aligned}$$

其中 $\boldsymbol{X}_{-i}$ 和 $\boldsymbol{Y}_{-i}$ 分别表示去掉第 i 个观测值的数据矩阵。

如果记 $\widetilde{\varepsilon}_i = Y_i - \widetilde{Y}_i$，其中 $\widetilde{Y}_i = \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_{(-i)}$，证明：

(a).

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}X_i\widetilde{\varepsilon}_i$$

(b).

$$\widetilde{\varepsilon}_i = (1 - h_{ii})^{-1}\widehat{\varepsilon}_i$$

第 4 讲 违背基本假设的处理

内容提要

- 异方差
- 自相关
- 异常值

4.1 引言

首先，我们回顾基本假设：Gauss-Markov 条件

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

- $E(\epsilon_i) = 0$，一般都满足。如果不满足，这一部分的效应将会叠加在常数项。

- $cov(\epsilon_i) = cov(\epsilon_j), i \neq j$ 是同方差假设（Homoskedasticity）。如果不满足，我们称为异方差（Heteroskedasticity）。
- $cov(\epsilon_i, \epsilon_j) = 0, i \neq j$ 是不相关假设。如果不满足，我们称为自相关。

我们知道模型显著性检验和系数显著性检验均是建立在同方差和不相关假设上。如果假设不成立，那么这两种检验全部失效，也就是虽然统计显著但实际不显著，或者说统计不显著但实际显著。换句话说，实际上自变量对因变量有影响，但统计检验不显著；或者实际上自变量对因变量没有有影响，但统计检验显著。所以，这次我们需要谨慎，寻找更加的检验方法（这一部分我们将在分位数回归进一步阐述）。

4.2 异方差性

4.2.1 原因和带来后果

例题 4.1 Engel **数据** Engel 数据包含 235 个观测值，2 个变量：

- income：家庭收入连续变量自变量
- fooexp：食物支出连续变量因变量

该数据主要用来研究家庭支出和家庭收入之间的关系，可以在 quantreg 中 engel 找到。

Engel 数代码

```
rm(list=ls())
library(quantreg)
data(engel)
head(engel)
attach(engel)
plot(income,foodexp,xlab="Household Income", ylab="Food Expenditure", cex=.5)
```

由于各户的收入、消费观念和习惯不同。低收入家庭购买的差异比较小，大多数购买生活必需品。高收入家庭购买差异较大：房子、汽车和股票等。低收入的家庭购买差异性比较小，高收入的家庭购买行为差异很大。导致消费模型的随机项 ε 具有不同的方差。异方差出现的原因可能是某一因素或某些因素随着自变量观察值的变化而对因变量产生不同的影响。

例题 4.2 模拟实验 我们考虑如下模型：

$$y_i = 1.5x_{1i} + 2x_{2i} + 1.5x_{3i} + x_{1i} * \varepsilon_i, i = 1, \ldots, n$$

其中

- $E(\varepsilon_i) = 0$
- $Cov(\varepsilon_i) = i, i = 1, \ldots, 10$，每 10 一组。

这里 ε_i 显然是异方差。我们设定样本 $n = 20$。

输出结果显示，$\hat{\beta}_1$ 是负值，显然严重错误。另外，除了 $\hat{\beta}_2$ 的估计比较准确外，其他参数估计均不准确，并且显著性有误。这说明：当存在异方差时，普通最小二乘估计存在以下问题：

(1) 参数估计值虽是无偏的，但不是最小方差线性无偏估计；

(2) 参数的显著性检验失效；

(3) 回归方程的应用效果极不理想，因为估计参数不准确。

笔记

- 异方差出现，但 $E(\varepsilon|X)=0$。从而

$$\begin{aligned}E(\widehat{\boldsymbol{\beta}}) &= E[E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})] \\ &= E\{E[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}|\boldsymbol{X}]\} \\ &= E\{E[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|\boldsymbol{X}]\} \\ &= E\{E[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{X}] + E[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}|\boldsymbol{X}]\} \\ &= E\{E[\boldsymbol{\beta}|\boldsymbol{X}]\} + E\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top E(\boldsymbol{\varepsilon}|\boldsymbol{X})\} \\ &= \boldsymbol{\beta}\end{aligned}$$

  我们可以得到参数估计值虽是无偏的。根据大数定律，只要模型正确设定。最小二乘估计仍具有一致性。

- 下面我们考虑其方差，

$$Var(\widehat{\boldsymbol{\beta}}) = Var[E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})] + E[Var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})]$$

从前面无偏性，我们可以得到 $Var[E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})] = 0$.

$$
\begin{aligned}
&E[Var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})] \\
&= E\{Var((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}|\boldsymbol{X})\} \\
&= E\{Var((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|\boldsymbol{X})\} \\
&= E\{Var((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}|\boldsymbol{X})\} \\
&= E\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top Var(\boldsymbol{\varepsilon})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}|\boldsymbol{X})\}
\end{aligned}
$$

如果 $Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}_n$，则 $Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\sigma^2$

如果 $Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{D} = diag\{\sigma_1^2, \ldots, \sigma_n^2\}$，从而

$$Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$

此时，$\sigma^2\boldsymbol{I}_n$ 不在是最小二乘估计的方差，因此，不能继续利用其构造置信区间和假设检验。

笔记 异方差带来的影响可能随着样本量的增大而减小。大家可以尝试 $n = 1000, 2000, 10000, 20000$ 看看输出结果。

模拟实验代码

```
> rm(list=ls())
> library(MASS)
```

```
> p <- 3; n <- 20
> mu <- rep(0, p)
> sigmu <- diag(rep(1, p))
> x <- MASS::mvrnorm(n=n, mu=mu, Sigma = sigmu)
> sigma <- rep(1:10, length=n)
> e <- rnorm(n=n, mean=0, sd=sigma)
> beta0 <- c(1.5, 2, 1.5)
> y <- x %*% beta0 + x[, 1]* e
> plot(x[, 1], y)
> fit <- lm(y~x-1)
> summary(fit)

Call:
lm(formula = y ~ x - 1)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-9.8871 -1.6389 -0.0347  1.0915  5.9751

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1  -0.6889     0.8595  -0.801    0.434
x2   2.0667     0.7131   2.898    0.010 *
x3   0.9606     1.1402   0.842    0.411
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.576 on 17 degrees of freedom
Multiple R-squared:  0.3577,    Adjusted R-squared:  0.2444
F-statistic: 3.156 on 3 and 17 DF,  p-value: 0.05183
```

4.2.2 异方差性的检验

常见的方法主要有残差分析法、等级相关系数法和 Breusch–Pagan 检验法等。

4.2.2.1 残差图分析法

4.2.2.2 等级相关系数法

等级相关系数检验法又称斯皮尔曼 (Spearman) 检验，是一种应用较广泛的方法。这种检验方法既可用于大样本，也可用于小样本。进行等级相关系数检验通常有三个步骤。

(1) 作 y 关于 x 的普通最小二乘回归，求出 ϵ_i 的估计值，即 e_i 的值。

(2) 取 e_i 的绝对值，分别把 x_i 和 $|e_i|$ 按递增（或递减）的次序分成等级，按下式计算出等级相关系数

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2$$

式中 n 为样本量；d_i 为对应于 x_i 和 $|e_i|$ 的等级差数。

(3) 做等级相关系数的显著性检验。在 $n > 8$ 的情况下，用下式对样本等级相关系数 r_s 进行

t 检验。检验统计量为:

$$t = \frac{\sqrt{n-2}r_s}{\sqrt{1-r_s^2}}$$

- 如果 $|t| \leq t_{\alpha/2}(n-2)$ 可认为异方差性问题不存在;
- 如果 $|t| > t_{\alpha/2}(n-2)$, 说明 x_i 和 $|e_i|$ 之间存在系统关系，异方差性问题存在。

4.2.2.3 Breusch–Pagan 检验法

Breusch–Pagan 检验是由 Breusch 核 Pagan(1979) 提出，其认为异方差有自变量与误差项相关造成的。由于误差项未知，所以使用拟合值代替，即假设

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{1i} + \cdots + x_{pi} + u_i \tag{4.1}$$

这个假设实际上认为 $\hat{\varepsilon}$ 与自变量有关。这里可以取全部的自变量，也可以部分自变量。如果模型不显著，即 $\alpha_1 = \cdots = \alpha_p = 0$，则同方差。问题转化为检验

$$\alpha_1 = \cdots = \alpha_p$$

这个检验实际上可以采用线性模型 (4.1) 的 F 检验实现。Breusch 核 Pagan(1979) 利用模型 (4.1) 的决定系数 R^2 构造了一个新的统计量 nR^2，并且证明其渐近服从自由度为 p 的卡方分

布。Koenker(1981) 提出了改进的统计量。R 的 lmtest 中的函数 bptest 可以实现。

4.2.2.4 Goldfeld-Quandt 检验法

Goldfeld-Quandt 是由 Goldfeld 和 Quandt (1965)，其将数据分成两部分，分别进行建模，进而得到两部分的 SSE_1 和 SSE_2，构造

$$F = \frac{SSE_1/df_1}{SSE_2/df_2}.$$

如果两者差异很大，所以存在异方差。其原假设是不存在异方差。R 的 lmtest 中的函数 qqtest 可以实现。

这四种方法各有优劣。残差图分析法简单，但比较主观。等级相关系数法需要残差与不同的自变量分别进行检验。Breusch–Pagan 检验法假设方差与自变量是线性关系，如果是非线性关系，导致检验效果不高。Goldfeld-Quandt 检验法将数据分成两组。如果两组组内部差异比较大，但组间差异比较小，也将导致方法效果下降。

4.2.2.5 Engel 数据分析

我们现在对 Engel 数据进行建模，并且进行异方差检验。

- 等级相关系数法拒绝原假设，存在异方差。
- Breusch–Pagan 检验的下 $p-value = 7.3110^{-6}$，所以拒绝原假设，认为存在异方差。
- Goldfeld-Quandt 检验没有拒绝原假设，不存在异方差。

我们可以看出不用的方法结论不一样。假设检验的原则是有一个拒绝便拒绝。所以异方差存在。

笔记 大家尝试模拟数据建模的异方差检验。思考检验的结果。

Engel 数据异方差检验

```
> fit1 <- lm(foodexp~income)
> cor.test(x=income, y =abs(fit1$residuals), method = "spearman")

        Spearman's rank correlation rho

data:  income and abs(fit1$residuals)
S = 1350113, p-value = 2.687e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
```

```
      rho
0.3757973

> lmtest::bgtest(fit1)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  fit1
LM test = 20.11, df = 1, p-value = 7.31e-06
> lmtest::gqtest(fit1)

        Goldfeld-Quandt test

data:  fit1
GQ = 1.1575, df1 = 116, df2 = 115, p-value = 0.2167
alternative hypothesis: variance increases from segment 1 to 2
```

4.2.3 消除异方差的方法

消除异方差性的方法通常有：

- 加权最小二乘法，
- Box-Cox 变换法
- 方差稳定性变换法

这里我们主要介绍加权最小二乘。加权最小二乘法 (Weighted Least Square, 简记为 WLS) 是一种最常用的消除异方差性的方法。

4.2.3.1 一元加权最小二乘估计

一元线性回归普通最小二乘法的残差平方和为：

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-E(y_i))^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

一元线性回归的加权最小二乘的离差平方和为：

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}\omega_i(y_i-E(y_i))^2=\sum_{i=1}^{n}\omega_i(y_i-\beta_0-\beta_1x_i)^2$$

加权最小二乘估计为

$$\begin{cases}\hat{\beta}_{0\omega}=\bar{y}_{\omega}-\hat{\beta}_{1\omega}\bar{x}_{\omega}\\ \hat{\beta}_{1\omega}=\dfrac{\sum_{i=1}^{n}\omega_i(x_i-\bar{x}_{\omega})(y_i-\bar{y}_{\omega})}{\sum_{i=1}^{n}\omega_i(x_i-\bar{x}_{\omega})^2}\end{cases}$$

其中 $\bar{x}_{\omega}=\frac{1}{\sum\omega_i}\omega_i x_i$ 为自变量的加权平均，$\bar{y}_{\omega}=\frac{1}{\sum\omega_i}\omega_i y_i$ 为自变量的加权平均。

为了消除异方差的影响，观测值的权数应该是观测值误差项方差的倒数，即

$$\omega_i=\frac{1}{\sigma_i^2}$$

σ_i^2 为第 i 个观测值误差项方差。误差项方差较大的观测值接受较小的权数；误差项方差较小的观测值接受较大的方差。

在社会经济研究中，经常会遇到误差项方差与 x 的幂函数 x^m 成比例，其中，m 为待定未知参数

$$\omega_i=\frac{1}{x_i^m}$$

4.2.3.2 多元加权最小二乘估计

当误差项 ϵ_i 存在异方差性时，对于一般的多元线性回归模型，加权离差平方和为

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

记

$$\boldsymbol{W} = \begin{pmatrix} \omega_1 & & & \vdots \\ & \omega_2 & & \\ & & \ddots & \\ \vdots & & & \omega_n \end{pmatrix}$$

加权最小二乘估计的矩阵表达式为

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

通常取权函数 $\boldsymbol{W}$ 为某个自变量 $x_j, j = 1, 2, \ldots, p)$ 的幂函数，即 $\boldsymbol{W} = x_j^m$

★ 在 $x_1, x_2, \cdots, x_p$ 这 p 个自变量中取哪一个？

这只需计算每个自变量 x_j 与普通残差的等级相关系数，选取等级相关系数最大的自变量构造权函数。

4.2.4 Engel 数据分析

我们继续 Engel 数据分析。我们使用决定系数选择 m。从结果来看，加权最小二乘将决定系数从 0.8304 提高到 0.8631，但从检验结果来看，仍然没有消去异方差。

- 加权最小二乘是以牺牲大方差项的拟合效果为代价改善了小方差项的拟合效果，这也并不总是研究者所需要的。
- 在社会经济现象中，通常变量取值大时方差也大，在以经济总量为研究目标时，更关心的是变量取值大的项，而普通最小二乘恰好能满足这个要求。
- 在时间序列数据建模时，近期的经济数据往往数值偏大，早期的数据数值会偏小，表现出异方差。
- 对这样的数据使用加权最小二乘，会对早期的数据拟合的更好，而近期的数据拟合效果变差。
- 在一些特定场合下，即使数据存在异方差，也仍然可以选择使用普通最小二乘估计。

Engel 数据加权最小二乘

```
> # select m begin
> m_vec <- seq(-4, 4, 0.1)
> m_len <- length(m_vec)
> R2_full <- NULL
> for(k in 1:m_len){
+   weig  <- income^(m_vec[k])
+   fit_weig <- lm(foodexp~income, weights = weig)
+   fitt <- summary(fit_weig)
+   R2_full[k] <- fitt$r.squared
+ }
> index_opt <- which.max(R2_full)
> m_opt <- m_vec[index_opt]
> weig_opt  <- income^(m_opt)
> fit_opt <- lm(foodexp~income, weights = weig_opt)
> summary(fit_opt)

```

```
Call:
lm(formula = foodexp ~ income, weights = weig_opt)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.38391 -0.09268  0.00610  0.09181  0.36149

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.3070    11.3283    6.03 6.38e-09 ***
income        0.5712     0.0149   38.33  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1235 on 233 degrees of freedom
Multiple R-squared:  0.8631,    Adjusted R-squared:  0.8625
```

```
F-statistic:  1469 on 1 and 233 DF,  p-value: < 2.2e-16

> summary(fit1)

Call:
lm(formula = foodexp ~ income)

Residuals:
    Min      1Q  Median      3Q     Max
-725.70  -60.24   -4.32   53.41  515.77

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 147.47539   15.95708   9.242   <2e-16 ***
income        0.48518    0.01437  33.772   <2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 114.1 on 233 degrees of freedom
Multiple R-squared:  0.8304,    Adjusted R-squared:  0.8296
F-statistic:  1141 on 1 and 233 DF,  p-value: < 2.2e-16

> cor.test(x=income, y =abs(fit_opt$residuals), method = "spearman")

        Spearman's rank correlation rho

data:  income and abs(fit_opt$residuals)
S = 1121381, p-value = 4.799e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4815476
```

```
> lmtest::bgtest(fit_opt)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  fit_opt
LM test = 20.11, df = 1, p-value = 7.31e-06

> lmtest::gqtest(fit_opt)

        Goldfeld-Quandt test

data:  fit_opt
GQ = 1.1575, df1 = 116, df2 = 115, p-value = 0.2167
alternative hypothesis: variance increases from segment 1 to 2
```

4.3 自相关性问题及其处理

4.3.1

如果一个回归模型的随机误差项 $cov(\epsilon_i, \epsilon_j) \neq 0$ 则称随机误差项之间存在着自相关现象。

这里的自相关现象不是指两个或两个以上的变量之间的相关，而指的是一个变量前后期数值之间存在的相关关系。

自相关性产生的背景和原因

(1) 遗漏关键变量时会产生序列的自相关性。

(2) 经济变量的滞后性会给序列带来自相关性。

(3) 采用错误的回归函数形式也可能引起自相关性。

(4) 蛛网现象 (Cobweb phenomenon) 可能带来序列的自相关性。

(5) 因对数据加工整理而导致误差项之间产生自相关性。

(1) 参数的估计值不再具有最小方差线性无偏性。

(2) 均方误差 MSE 可能严重低估误差项的方差。

(3) 容易导致对 t 值评价过高，常用的 F 检验和 t 检验失效。如果忽视这一点，可能导致得出回归参数统计检验为显著，但实际上并不显著的严重错误结论。

(4) 当存在序列相关时，仍然是 β 的无偏估计量，但在任一特定的样本中，可能严重歪曲 β 的真实情况，即最小二乘估计量对抽样波动变得非常敏感

(5) 如果不加处理地运用普通最小二乘法估计模型参数，用此模型进行预测和结构分析将会带来较大的方差甚至错误的解释。

4.3.2

- 图示检验法

(1) 绘制 (e_t, e_{t-1}) 的散点图。

- 图示检验法

(2) 按照时间顺序绘制回归残差项 e_t 的图形

4.3.3

- 自相关系数法

误差序列 $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ 的自相关系数定义为：

$$\rho = \frac{\sum_{t=2}^{n} \epsilon_t \epsilon_{t-1}}{\sqrt{\sum_{t=2}^{n} \epsilon_t^2} \sqrt{\sum_{t=2}^{n} \epsilon_{t-1}^2}}$$

ρ 的取值范围是 $[-1, 1]$，当 ρ 接近 1 时，表明序列误差存在正相关，当 ρ 接近 -1 时，表示序列误差存在负相关。

在实际应用中，误差序列 $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ 的值是未知的，需要用其估计值 e_i 代替，得到自相关系数的估计值为

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sqrt{\sum_{t=2}^{n} e_t^2} \sqrt{\sum_{t=2}^{n} e_{t-1}^2}}$$

4.3.4

- DW 检验

DW 检验是 J.Durbin 和 G.S.Watson 于 1951 年提出的一种适用于小样本的一种检验方法。

DW 检验只能用于检验随机扰动项具有一阶自回归形式的序列相关问题。

这种检验方法是建立计量经济学模型中最常用的方法，一般的计算机软件都可自动产生出 D.W 值。

随机误差项的一阶自回归形式为

$$\epsilon_t = \rho\epsilon_{t-1} + \mu_t$$

为了检验序列的相关性，构造的假设是

$$H_0: \quad \rho = 0$$

- DW 检验

定义 DW 统计量为：

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2}$$

如果认为 $\sum_{t=2}^{n} e_t^2$ 与 $\sum_{t=2}^{n} e_{t-1}^2$ 近似相等

$$DW = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2} \approx 2\left[1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}\right]$$

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sqrt{\sum_{t=2}^{n} e_t^2}\sqrt{\sum_{t=2}^{n} e_{t-1}^2}} \approx \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}$$

因此

$$DW \approx 2(1-\hat{\rho})$$

根据样本容量 n 和解释变量的数目 k(这里包括常数项), 查 DW 分布表, 得临界值 d_L 和 d_U, 然后依下列准则考察计算得到的 DW 值, 以决定模型的自相关状态:

D.W 检验尽管有着广泛的应用, 但也有明显的缺点和局限性。

- DW 检验有一个不能确定的区域, 一旦 DW 值落在这个区域, 就无法判断。这时, 只有增大样本容量或选取其他方法。
- DW 统计量的上、下界表要求 $n > 15$, 这是因为样本如果再小, 利用残差就很难对自相关的存在性作出比较正确的诊断。
- DW 检验不适应随机项具有高阶序列相关的检验。

4.3.5

- 迭代法

以一元线性回归模型为例，设一元线性回归模型的误差项存在一阶自相关

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

$$\epsilon_t = \rho \epsilon_{t-1} + \mu_t$$

$$\begin{cases} E(\mu_t) = 0, \quad t = 1, 2, \ldots, n \\ cov(\mu_t, \mu_s) = \begin{cases} \sigma^2, t = s \\ 0, t \neq s \end{cases} \quad t, s = 1, 2, \ldots, n \end{cases}$$

根据一元线性回归模型 $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$，有

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}$$

变形后有

$$(y_t - \rho y_{t-1}) = (\beta_0 - \rho\beta_0) + \beta_1 (x_t - \rho x_{t-1}) + (\epsilon_t - \epsilon_{t-1})$$

令：

$$y_t' = y_t - \rho y_{t-1}$$

$$x_t' = x_t - \rho x_{t-1}$$

$$\beta_0' = \beta_0(1-\rho)$$

$$\beta_1' = \beta_1$$

得到有随机独立误差项，满足线性回归基本假设的

$$y_t' = \beta_0' + \beta_1' x_t' + \mu_t \qquad (*)$$

其中自相关系数 ρ 用公式 $\hat{\rho} \approx 1 - \frac{1}{2}DW$ 估计。

用变换因变量与变换自变量作普通最小二乘回归。如果误差项确实是一阶自相关，通过以上变换，回归模型已经消除自相关。

实际问题中，有时误差项并不是简单的一阶自相关，而是更复杂的自相关形式，(*) 式的误差项 u_t 可能仍然存在自相关，这就需要进一步对 (*) 式的误差项 u_t 做 DW 检验，以判断 u_t 是否存在自相关，如果检验表明误差项 u_t 不存在自相关，迭代法到此结束。如果检验表明误差项 u_t 存在自相关，那末对回归模型 (*) 式重复用迭代法，这个过程可能要重复几次，直至最终消除误差项自相关。这种迭代消除自相关的过程正是迭代法名称的由来。

- 差分法

一阶差分法通常适用于原模型存在较高程度的一阶自相关的情况。在迭代法中，当 ρ

$$(y_t - \rho y_{t-1}) = (\beta_0 - \rho\beta_0) + \beta_1(x_t - \rho x_{t-1}) + (\epsilon_t - \epsilon_{t-1})$$

为：

$$(y_t - \rho y_{t-1}) = \beta_1(x_t - x_{t-1}) + (\epsilon_t - \epsilon_{t-1})$$

以 $\Delta y_t = y_t - y_{t-1}$，$\Delta x_t = x_t - x_{t-1}$，得到

$$\Delta y_t = \beta_1 \Delta x_t + \mu_t$$

上式是不带有常数项的回归方程

$$\hat{\beta}_1 = \frac{\sum_{t=2}^{n} \Delta y_t \Delta x_t}{\sum_{t=2}^{n} \Delta x_t^2}$$

一阶差分法的应用条件是自相关系数 $\rho = 1$，在实际应用中，ρ 接近 1 时我们就采用差分法而不用迭代法，这有两个原因。

第一，迭代法需要用样本估计自相关系数 ρ，对 ρ 的估计误差会影响迭代法的使用效率；

第二，差分法比迭代法简单，人们在建立时序数据的回归模型时，更习惯于用差分法。

但是完全的 $\rho = 1$ 情况几乎是见不到的，实际应用时 ρ 较大就行！

4.4 BOX-COX 变换

4.4.1

BOX-COX 变换是由博克斯（Box）与考克斯（Cox）在 1964 年提出的一种应用非常广泛的变换，它是对因变量 y 做如下变换：

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \ln y, \lambda = 0 \end{cases}$$

其中，λ 为待定参数。此变换要求 y 的各分量都大于 0。否则可用下面推广的 BOX-COX 变换

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+a)^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \ln(y+a), \lambda = 0 \end{cases}$$

即先对 y 做平移，使得 $y + a$ 的各个分量都大于 0 后再做 Box-Cox 变换。

对于不同的 λ，所做的变换也不同，所以这是一个变换族。它包含一些常用的变换，如对数变换 ($\lambda = 0$)，平方根变换 ($\lambda = 1/2$) 和倒数变换 ($\lambda = -1$)。

寻找合适的 λ，使得变换后

$$\boldsymbol{y}^{(\lambda)}=\begin{pmatrix}y_1^{(\lambda)}\\ y_2^{(\lambda)}\\ \vdots\\ y_n^{(\lambda)}\end{pmatrix}\sim N_n(\boldsymbol{X\beta},\sigma^2\boldsymbol{I})$$

从而符合线性回归模型的各项假设：误差分量等方差、不相关等。

事实上，BOX-COX 变换不仅可以处理异方差性问题，还能处理自相关、误差非正态、回归函数非线性等情况。

经过计算可得 λ 的最大似然估计（参见参考文献 [2]）

$$L_{max}(\lambda)=(2\pi e\hat{\sigma}_\lambda^2)^{-\frac{n}{2}}|\boldsymbol{J}|$$

式中，$\hat{\sigma}_\lambda^2=\frac{1}{n}SSE(\lambda,y^{(\lambda)})$，$|\boldsymbol{J}|=\prod_{i=1}^n|\frac{dy_i^{(\lambda)}}{dy_i}|=\prod_{i=1}^n y_i^{(\lambda-1)}$

令 $z^{(\lambda)}=\frac{y^{(\lambda)}}{|\boldsymbol{J}|}$，对 L_{max} 取对数并略去与 λ 无关的常数项，可得

$$\ln L_{max}(\lambda)=-\frac{n}{2}\ln SSE(\lambda,z^{(\lambda)})$$

为找出 λ，使得 $\ln L_{max}(\lambda)$ 达到最大，只需使 $SSE(\lambda, z^{(\lambda)})$ 达到最小即可。它的解析解比较难找，通常是给出一系列 λ 的值，计算对应的 $SSE(\lambda, z^{(\lambda)})$，取使得 $SSE(\lambda, z^{(\lambda)})$ 达到最小的 λ 即可。

- BOX-COX 变换是一个幂变换族，其中当变换参数 $\lambda = 0$ 时成为对数变换，而对数变换则是比幂变换应用更广泛的变换，很多场合都可以首先尝试对数据作对数变换。
- 从概率分布的角度看，当数据本身服从对数正态分布时，对数据取对数变换后就服从正态分布。对数正态分布是右偏分布，有厚重的右尾。
- 从数据看，如果数据中一些数值很大，但是小数值的数据更密集，个数也更多，大数值的数据较较疏松，个数较少，这样的数据很可能服从对数正态分布，可以尝试对数变换。
- 对回归分析问题，如果只对因变量作对数变换，就是 BOX-COX 变换 $\lambda = 0$ 时的特例。也可以考虑只对自变量作对数变换，或者同时对因变量和对自变量作对数变换。

4.5 异常值与强影响点

异常值分为两种情况：

- 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
- 补充材料：CH4 Regression Analysis By Example 5th

4.5.1

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。

4.5.2

- 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

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

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

Consider a random sample $X_1, X_2, \ldots, X_n \sim Unif(0, \theta)$.

(i) Find the estimator for θ through MoM,denoted by $\hat{\theta}_{MM}$.

(ii) Find the MLE $\hat{\theta}_{MLM}$.

(iii) What are the expetation and variance of $\hat{\theta}_{MM}$ and $\hat{\theta}_{MLM}$? Which estimator is better?

4.6 参考文献

- Breusch, T. S., Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, 1287–1294.
- Goldfeld S.M., Quandt R.E. (1965), Some Tests for Homoskedasticity. Journal of the American Statistical Association 60, 539–547
- Koenker R.(1981), A Note on Studentizing a Test for Heteroscedasticity. Journal of Econometrics 17, 107–112.

第 4 讲 练习

1. diabetes 数据进行异方差、自相关和异常值进行分析 4

图 4.1: Engel 数据

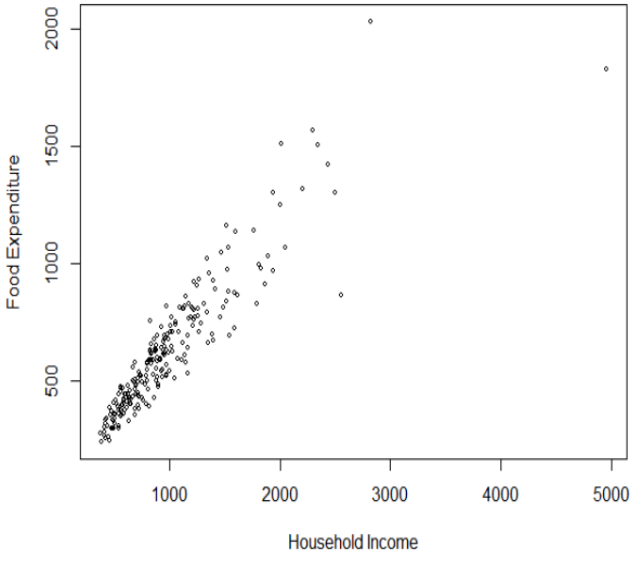

Food Expenditure
500
1000
1500
2000
1000
2000
3000
4000
5000
Household Income

图 4.2: 无异方差

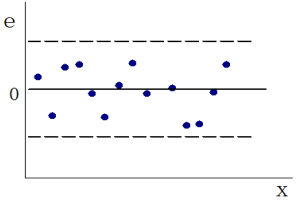
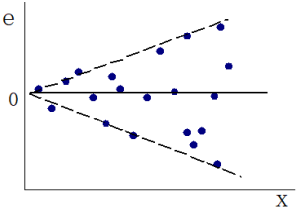
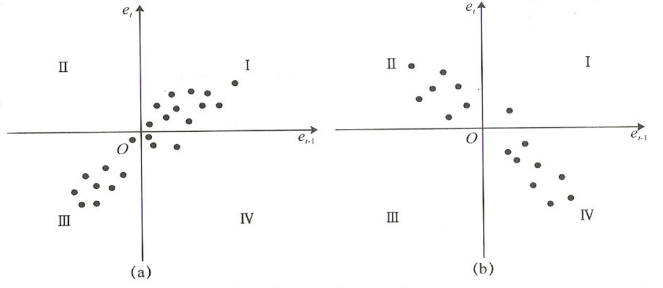
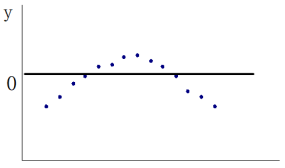


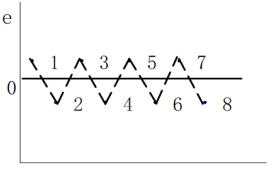
图 4.3: 存在异方差







(c)

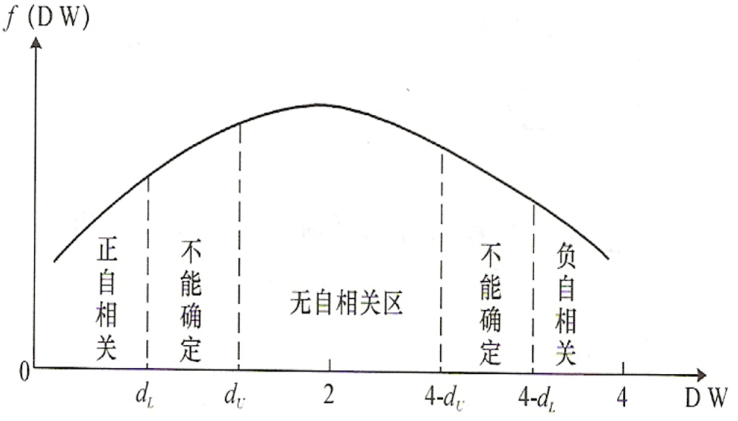


(d)

图 4.4: DW 值与 $\hat{\rho}$ 对应关系

$\hat{\rho}$	D.W	误差项的自相关性
-1	4	完全负自相关
(-1，0)	(2，4)	负自相关
0	2	无自相关
(0，1)	(0，2)	正自相关
1	0	完全正自相关

$0 \leqslant D.W \leqslant d_L$,	误差项 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 间存在正相关;
$d_L < D.W \leqslant d_U$,	不能判定是否有自相关;
$d_U < D.W < 4-d_U$,	误差项 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 间无自相关;
$4-d_U \leqslant D.W < 4-d_L$,	不能判定是否有自相关;
$4-d_L \leqslant D.W \leqslant 4$,	误差项 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 间存在负相关。


f (D W)
正自相关
不能确定
无自相关区
不能确定
负自相关
0
d_L
d_U
2
4-d_U
4-d_L
4
D W

异常值原因	异常值消除方法
1. 数据登记误差，存在抄写或录入的错误	重新核实数据
2. 数据测量误差	重新测量数据
3. 数据随机误差	删除或重新观测异常值数据
4. 缺少重要自变量	增加必要的自变量
5. 缺少观测数据	增加观测数据，适当扩大自变量取值范围
6. 存在异方差	采用加权线性回归
7. 模型选用错误，线性模型不适用	改用非线性回归模型

第 5 讲 多重共线性的情形及其处理

内容提要

- 多重共线性
- 岭回归
- 变量选择

5.1 diabetes 数据续

前面分析，我们得到很多自变量不显著。这一讲我们将进一步探索。

5.2 多重共线性

5.2.1 表述定义

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

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

  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。

5.2.2 多重共线性产生背景和原因

- 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。

5.2.3 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%。

表 6.1

r_{12}	0.0	0.2	0.50	0.70	0.80	0.90	0.95	0.99	1.00
$var(\hat{\beta}_1)$	1.0	1.04	1.33	1.96	2.78	5.26	10.26	50.25	∞

5.2.4 多重共线性的诊断

常见的诊断方法主要有方差扩大因子和特征根判定法等。

5.2.4.1 方差扩大因子法

对自变量做中心标准化，则 $\boldsymbol{X}^{*\top}\boldsymbol{X}^{*\top} = (r_{ij})$ 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)。经过简单计算，

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

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。R 包 car 的 vif 函数可以实现。

5.2.4.2 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

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_px_{ip}\approx 0$$

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；
  - $10 \leq k < 100$ 时，认为 $\boldsymbol{X}$ 存在较强的多重共线性；
  - 当 $k \geq 100$ 时，则认为存在严重的多重共线性。

5.2.4.3 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。

笔记

- 当解释变量之间的简单相关系数很大时，可以断定自变量间存在着严重的多重共线性；
- 但是一个回归方程存在严重的多重共线性时，解释变量之间的简单相关系数不一定很大。
  - 例如假定 3 个自变量之间有完全确定的关系

$$x_1 = x_2 + x_3$$

再假定 x_2 与 x_3 的简单相关系数 $r_{23} = -0.5$，x_2 与 x_3 的离差平方和 $L_{22} = L_{33} = 1$，此时

$$L_{23} = r_{23}\sqrt{L_{22}L_{33}} = -0.5$$

$$
\begin{aligned}
L_{11} &= \sum (x_1 - \bar{x}_1)^2 = \sum (x_2 + x_3 - (\bar{x}_2 + \bar{x}_3))^2 \\
&= \sum ((x_2 - \bar{x}_2) + (x_3 - \bar{x}_3))^2 \\
&= \sum (x_2 - \bar{x}_2)^2 + \sum (x_3 - \bar{x}_3)^2 + 2 \sum (x_2 - \bar{x}_2)(x_3 - \bar{x}_3) \\
&= 1 + 1 + 2 \times (-0.5) = 1
\end{aligned}
$$

$$
\begin{aligned}
L_{12} &= \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \\
&= \sum (x_2 + x_3 - (\bar{x}_2 + \bar{x}_3))(x_2 - \bar{x}_2) \\
&= \sum (x_2 - \bar{x}_2)^2 + \sum (x_3 - \bar{x}_3)(x_2 - \bar{x}_2) \\
&= L_{22} + L_{23} = 1 - 0.5 = 0.5
\end{aligned}
$$

因而 $r_{12} = L_{12}/\sqrt{L_{11}L_{22}} = 0.5$

同理 $r_{13} = 0.4$

由此看到，当回归方程中的自变量数目超过 2 时，并不能由自变量间的简单相关系数不高，就断定它们不存在多重共线性。

5.3 消除多重共线性的方法

- 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。
- 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。

笔记 增大样本量一般不易做到。大家可以试试模拟实验。

5.4 岭回归

5.4.1 模型表达

当自变量间存在复共线性时，回归系数估计的方差就很大，估计值就很不稳定。在具体取值上与真实值有较大的偏差，有时甚至会出现与实际意义不服的正负号。

岭回归 (Ridge Regression) 提出的想法是很自然的。当自变量间存在复共线性时，$\boldsymbol{X}'\boldsymbol{X} \approx 0$。我们可以在 $\boldsymbol{X}'\boldsymbol{X}$ 加上一个正常数矩阵 $k\boldsymbol{I}, (k > 0)$，使得 $\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}$ 可逆。从而，我们构造估计量：

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y} \tag{5.1}$$

还一个角度，岭回归是求带有约束最小二乘：

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\top}\beta)^2 \tag{5.2}$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leq t \tag{5.3}$$

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

$$\min_{\beta}\left(\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^{\top}\beta)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right) \tag{5.4}$$

(6.8) 求解便是 (5.1)。

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。

5.4.2 估计量性质

性质 1 $\hat{\boldsymbol{\beta}}(k)$ 是回归参数 $\boldsymbol{\beta}$ 的有偏估计。

证明：

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

显然只有当 $k=0$ 时，$E(\hat{\boldsymbol{\beta}}(0))=\boldsymbol{\beta}$；当 $k\neq 0$，$\hat{\boldsymbol{\beta}}(k)$ 是回归参数 $\boldsymbol{\beta}$ 的有偏估计。要特别强调的是 $\hat{\boldsymbol{\beta}}(k)$ 不再是 $\boldsymbol{\beta}$ 的无偏估计，有偏性是岭回归估计的一个重要特性。

性质 2 在认为岭参数 k 是 $\boldsymbol{y}$ 的无关的常数时，

$\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，也是 $\boldsymbol{y}$ 的线性函数。

证明：

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

所以，岭估计 $\hat{\boldsymbol{\beta}}(k)$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，根据定义式 $\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 知 $\hat{\boldsymbol{\beta}}(k)$ 也是 $\boldsymbol{y}$ 的线性函数。

- 这里需要注意的是，在实际应用中，由于岭参数 k 总是要通过数据来确定，因而 k 也依赖于 $\boldsymbol{y}$，
- 本质上来说，$\hat{\boldsymbol{\beta}}(k)$ 并非 $\hat{\boldsymbol{\beta}}$ 的线性变换，也不是 $\boldsymbol{y}$ 的线性函数。

性质 3 对任意的 $k>0$，$||\hat{\boldsymbol{\beta}}||\neq 0$，总有

$$||\hat{\boldsymbol{\beta}}(k)||<||\hat{\boldsymbol{\beta}}||$$

这里 $||\cdot||$ 是向量的模，等于向量各分量的平方和。这个性质表明 $\hat{\boldsymbol{\beta}}(k)$ 可看成由 $\hat{\boldsymbol{\beta}}$ 进行某种

向原点的压缩。从 $\hat{\boldsymbol{\beta}}(k)$ 的表达式可以看到，当 $k \to \infty$ 时，$\hat{\boldsymbol{\beta}}(k) \to \mathbf{0}$，即 $\hat{\boldsymbol{\beta}}(k)$ 化为零向量。

性质 4 以 MSE 表示估计向量的均方误差，则存在 $k > 0$，使得

$$MSE[\hat{\boldsymbol{\beta}}(k)] < MSE[\hat{\boldsymbol{\beta}}]$$

即

$$\sum_{j=1}^{p} E[\hat{\beta}_j(k) - \beta_j]^2 < \sum_{j=1}^{p} D(\hat{\beta}_j)$$

k 的选择的选择，我们将在下一节进行。

5.5 自变量选择的准备

方法介绍前，我们先看看选择全模型和选模型，这有利于理解变量选择的原因。

5.5.1 全模型和选模型

- 全模型　　设研究某一实际问题涉及的对因变量有影响的因素共 m 个，由因变量 y 和 m 个自变量 $x_1, x_2, \cdots, x_m$ 构成的回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \epsilon \tag{5.5}$$

  称该模型为全回归模型

- 选模型　　如果从所有可选择的 m 个变量中挑选出 p 个，记为 $x_1, x_2, \cdots, x_p$，由所选的 p 个自变量组成的回归模型为

$$y = \beta_{0p} + \beta_{1p} x_1 + \beta_{2p} x_2 + \cdots + \beta_{pp} x_p + \epsilon \tag{5.6}$$

  称该模型为选模型。

模型选择不当会给参数估计和预测带来什么影响？下面我们将分别给予讨论。

为了方便，把全模型式 (5.5) 的参数向量 $\boldsymbol{\beta}$ 和 σ^2 的估计记为

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

$$\hat{\sigma}_m^2 = \frac{1}{n-m-1} SSE_m$$

把选模型式 (5.6) 的参数向量 $\boldsymbol{\beta}$ 和 σ^2 的估计记为

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

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

假设全模型式 (5.5) 与选模型式 (5.6) 不同，即要求 $p < m$，$\beta_{p+1}x_{p+1} + \cdots + \beta_m x_m$ 不恒为 0。在此条件下，当全模型正确而误用了选模型时，有以下性质

(1) 在 x_j 与 $x_{p+1}, \cdots, x_m$ 的相关系数不全为 0 时，选模型回归系数的最小二乘估计是全模型相应参数的有偏估计，即

$$E(\hat{\beta}_{jp}) = \beta_{jp} \neq \beta_j, j = 1, 2, \cdots, p$$

(2) 选模型的预测是有偏的。给定新的自变量值，

$\boldsymbol{x}_{0m} = (x_{o1}, x_{02}, \cdots, x_{0m})^\top$，因为新值为 $y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots, \beta_m x_{0m} + \epsilon_0$，用选模型的预测值 $\hat{y}_{0p} = \hat{\beta}_{0p} + \hat{\beta}_{01} x_{01} + \cdots + \hat{\beta}_{pp} x_{0p}$ 作为 y_0 的预测值是有偏的，即 $E(\hat{y}_{0p} - y_0) \neq 0$

(3) 选模型的参数估计有较小的方差。

选模型的最小二乘估计 $\hat{\boldsymbol{\beta}}_p = (\hat{\beta}_{0p}, \hat{\beta}_{1p}, \hat{\beta}_{2p}, \cdots, \hat{\beta}_{pp})^\top$，

全模型的最小二乘估计为 $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_{0m}, \hat{\beta}_{1m}, \hat{\beta}_{2m}, \cdots, \hat{\beta}_{mm})^\top$，

这一性质说明 $D(\hat{\beta}_{jp}) \leq D(\hat{\beta}_{jm}), j = 1, 2, \cdots, p$

(4) 选模型的预测残差有较小的方差。

选模型的预测残差为 $e_{0p} = \hat{y}_{0p} - y_0$，

全模型的预测残差为 $e_{0m}=\hat{y}_{0m}-y_0$，

其中 $y_0=\beta_0+\beta_1x_{01}+\beta_2x_{02}+\cdots+\beta_mx_{0m}+\epsilon$，则有 $D(e_{0p})\le D(e_{0m})$

(5) 记 $\boldsymbol{\beta}_{m-p}=(\beta_{p+1},\cdots,\beta_m)'$，用全模型对 $\boldsymbol{\beta}_{m-p}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}}_{m-p}=(\hat{\beta}_{p+1},\cdots,\hat{\beta}_m)'$，则在 $D(\boldsymbol{\beta}_{m-p})\ge\boldsymbol{\beta}_{m-p}\boldsymbol{\beta}'_{m-p}$ 的条件下，$E(e_{0p})^2=D(e_{0p})+(E(e_{0p}))^2\le D(e_{0m})$，即选模型预测的均方误差比全模型预测的方差小。

笔记

- 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。

5.5.2 几种常见准则

5.5.2.1 调整的 R^2

前面我们已经给出 R^2 和调整的 R^2，下面我们进一步讨论两者。

- 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。

这两种方法都有明显的不足，这是因为：

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

$$R^2_{p+1} \geq R^2_p$$

1 自由度调整复决定系数达到最大

- 调整的复决定系数为

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

- $R^2_a \leq R^2$，R^2_a 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。

5.5.2.2 AIC

- 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 $L(\boldsymbol{\theta})$，则 AIC 定义为：

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}_L) + 2p \tag{5.7}$$

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

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

将上式代入（5.7），这里似然函数中未知参数得个数为 $2p$，略去与 p 无关得常数，则回归模

型得 AIC 为

$$AIC = n\ln(SSE) + 2p \tag{5.8}$$

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。

笔记 虽然 (5.8) 基于正态分布假设得到，实际上即使不是正态分布仍然使用。

5.5.2.3 BIC

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

$$BIC = n\log(SSE) + \log(n)p \tag{5.9}$$

对比（5.8）和（5.9），我们发现 BIC 是 $\log(n)$，而 AIC 是 2。当 $n > 8$,$\log(n) > 2$，所以 BIC 对自变量数目的惩罚更大，得到的自变量个数更小。

5.5.2.4 C_p

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。

5.6 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

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

最优子集是利用上面的准则在所有自变量的组合中选择变量。

5.7 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。

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

5.7.0.1 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

5.7.0.2 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 $F_j^m \le F_\alpha(1, n-m-1)$，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
  $F_j^{m-1} \le F_\alpha(1, n-(m-1)-1)$，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。
- 续例 5.4

5.7.0.3 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 5 讲 练习

1. 假设 $X_1, X_2, \ldots, X_n \sim Unif(0, \theta)$.
   (i) 求矩估计 $\hat{\theta}_{MM}$.
   (ii) 求极大似然 $\hat{\theta}_{MLM}$.
   (iii) 计算两个估计的均值、方差、和均方误差。比较哪个估计更好。

第 6 讲 变量选择

内容提要

- LASSO
- 算法

假设回归模型是

$$Y = X^\top \beta + \varepsilon \tag{6.1}$$

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

$$E(\varepsilon | X = x) = 0, \tag{6.2}$$

模型 (6.1) 可以表示为：

$$E(Y | X = x) = x^\top \beta. \tag{6.3}$$

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

$$\min E(Y - x\beta)^2 \tag{6.4}$$

备注 假条条件 (6.2) 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 (6.4) 的样本实现值为

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\top}\beta)^2 \tag{6.5}$$

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 (6.2)，$\hat{\beta}_{ols}$ 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：

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \beta)^2 \tag{6.6}$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \le t \tag{6.7}$$

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

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\top} \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{6.8}$$

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。

6.1 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

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\top} \beta)^2 \tag{6.9}$$

$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{6.10}$$

上式等价于

$$\min_{\beta}\left(\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^{\top}\beta)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right) \tag{6.11}$$

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

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

其中 df 估计非零的参数的个数。$\lambda_{opt}=\arg\min_{\lambda_n}\mathrm{BIC}(\lambda_{\mathrm{n}})$.

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。

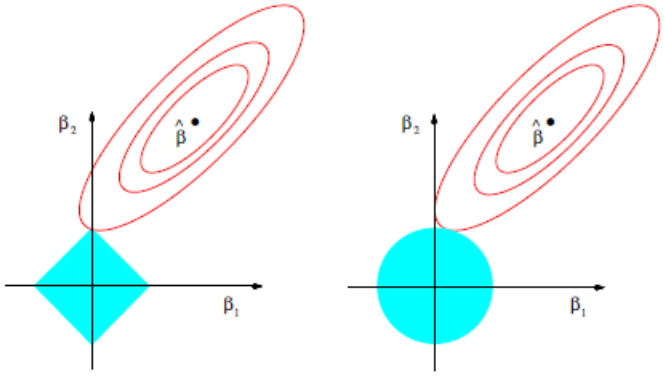


图 6.1: LAOO 和岭回归的几何解释；左边是 LASSO，右图是岭回归

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：

$$\min_{\beta}\left(\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^{\top}\beta)^2+\lambda\sum_{j=1}^{p}w_j|\beta_j|\right) \tag{6.12}$$

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：

$$\frac{1}{2}\|y-\boldsymbol{x}\beta\|^2+\lambda\sum_{j=1}^{p}p_{\lambda}(|\beta_j|) \tag{6.13}$$

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

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

这是因为:

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

模型 (6.13) 可以转化为

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

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 $\hat{\theta} = \text{sgn}(z)(|z| - \lambda)_+$
4. SCAD 估计量

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

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)：$\min_{\theta \neq 0}\{|\theta| + p'_\lambda(|\theta|)\} > 0$，则解具有稀疏性，即当 $|z| < \min_{\theta \neq 0}\{|\theta| + p'_\lambda(|\theta|)\}$ 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 6.1: 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	√	√	√

6.2 算法

6.2.1 二次逼近算法

关于惩罚函数求解的方法有很多。我们主要介绍局部二次逼近算法 (Local quadratic approximation,**Fan2001**)。该算法采用二次逼近罚函数。假设 β^* 比较接近最小值。

- 如果 β_j^* 非常接近 0，我们设定 $\hat{\beta}_j = 0$.
- 否则，我们使用二次逼近罚函数 $p_\lambda(|\beta_j|)$

由于，当 $\beta_j \neq 0$，且 $\beta_j^* \approx \beta_j$

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sgn}\beta_j \approx \frac{p'_\lambda(|\beta_j^*|)}{|\beta_j^*|}\beta_j.$$

换句话说：

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^*|) + \frac{1}{2}\frac{p'_\lambda(|\beta_j^*|)}{|\beta_j^*|}\left(\beta_j^2 - (\beta_j^*)^2\right) \tag{6.14}$$

上面的式子是 $\beta_j^2 - (\beta_j^*)^2$(不考虑 1/2)，不是 $|\beta_j| - |\beta_j^*|$。根据 Taloy 展开式这一项应该是绝对值的差，但作者使用了平方，因为平方可导。这也是称为二次逼近的原因。能不能用绝对值？当然可以。它的名字是局部线性逼近 (Local linear approximation,**Zou2008**)。二次逼近可导，线性逼近不可导；二次逼近的解不稀疏，而线性逼近的解稀疏。关于二次逼近的讨论，详见**Lee2016**。

将上面逼近代入一般的惩罚函数的表达式为：

$$\min\left\{(y - \boldsymbol{x}\beta)^\top(y - \boldsymbol{x}\beta) + \sum_{j=1}^{p}\left[p_\lambda(|\beta_j^*|) + \frac{1}{2}\frac{p'_\lambda(|\beta_j^*|)}{|\beta_j^*|}\left(\beta_j^2 - (\beta_j^*)^2\right)\right]\right\}$$

由于 β_j^* 是给定的，所以上面的式子进一步转化为：

$$\min \left\{ (y - \boldsymbol{x}\beta)^\top (y - \boldsymbol{x}\beta) + \sum_{j=1}^{p} \frac{1}{2} \frac{p'_\lambda(|\beta_j^*|)}{|\beta_j^*|} \beta_j^2 \right\} \tag{6.15}$$

为了记号方便，令 $u_i(\beta_j^*) = \frac{1}{2} \frac{p'_\lambda(|\beta_j^*|)}{|\beta_j^*|}$，则

$$\beta = (\boldsymbol{x}^\top \boldsymbol{x} + U(\beta^*))^{-1} \boldsymbol{x}^\top y \tag{6.16}$$

其中 $U(\beta^*) = \text{diag}\{u_1(\beta_j^*), \ldots, u_p(\beta_j^*)\}$.

具体算法如下：

1. 第一步：给定初始值 $\beta^{(0)}$,
2. 第二步：令 $\beta^{(m)} = \beta^{(0)}$ 根据 (6.16) 更新 $\beta^{(m+1)}$
3. 第三步：重复第二步，直到其收敛。

对于 Adaptive LASSO，而言，$u_i(\beta_j^*) = \lambda \frac{1}{2|\hat{\beta}_{ols,j}|} \frac{1}{|\beta_j^*|}$。我们进行下面模型模型看看算法效果如何。

在进行分析前，我们需要解决常数项的问题，变量惩罚不会对其对其进行惩罚。下面我们简单的说明一下。回归模型为

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

其经验的回归方程是：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_1 x_1 \tag{6.17}$$

经过中心化 $(\overline{x}_1, \ldots, \overline{x}_p, \overline{y})$ 的变量为 $(\widetilde{x}_1, \ldots, \widetilde{x}_p, \widetilde{y})$，则令

$$\widetilde{y} = \hat{\beta}_1^{\star} \widetilde{x}_1 + \cdots + \hat{\beta}_p^{\star} \widetilde{x}_p$$

进行得到：

$$y - \overline{y} = \hat{\beta}_1^{\star}(x_1 - \overline{x}_1) + \cdots + \hat{\beta}_p^{\star}(x_p - \overline{x}_p)$$

从而得到

$$y = (\overline{y} - \hat{\beta}_1^{\star}\overline{x}_1 - \cdots - \hat{\beta}_p^{\star}\overline{x}_p) + \hat{\beta}_1^{\star} x_1 + \cdots + \hat{\beta}_p^{\star} x_p \tag{6.18}$$

比较 (6.17) 和 (6.18)，我们可以得到：

$$\beta_0 = \overline{y} - \hat{\beta}_1^{\star}\overline{x}_1 - \cdots - \hat{\beta}_p^{\star}\overline{x}_p$$
$$\hat{\beta}_j = \hat{\beta}_j^{\star}, \quad j = 1, \ldots, p$$

下面是二次逼近的代码，我们可以修改阈值 (Threshold) 看看变量选择结果，默认 thre=1e-2，如果估计参数值的绝对值小于 0.01 设置为 0。从输出的结果来看，算法的效果还可以。

Adaptive LASSO 二次逼近代码

```
rm(list=ls())
library(MASS)

data1 <- function(n, p, sig0=0.25, rho=0.5){
  sigm <- sig0^abs(outer(1:(p),1:(p),"-"))
  muz <- rep(0, p)
  x <- mvrnorm(n = n, mu=muz, Sigma=sigm)
  e <- rnorm(n, 0, 1)
  beta <-c(3, 1.5, 0, 0, 2, rep(0, p-5))
  y <-  x %*% beta + rho * e
  return(list(y=y, x=x, beta=beta))
}

p <- 8
n <- 200
dat <- data1(n = n, p = p)
```

```
x <- dat$x
y <- dat$y
beta.ture <- dat$beta

#lam:lambda
#eps:
#itemax：maximum iteration time
#thre：  threshold
adalasso.my <- function(x, y, lam=0.1, eps=1e-5, itemax=1000, thre=1e-2, intercept=TRUE){
  p <- dim(x)[2]
  n <- dim(x)[1]
  x_colname <- colnames(x)
  beta_ols <- coef(lm(y~x-1))
  BB <- beta_ols
  if(is.null(x_colname)==TRUE){x_colname <- paste("x", 1:p, sep = "")}
  #
```

```
33  if(intercept==TRUE){
34    ##################################inclued intercept begin
35    x_mean <- apply(x, 2, mean)
36    y_mean <- mean(y)
37    x_c <- scale(x=x, center=TRUE, scale=FALSE)
38    y_c <- y - mean(y)
39    txx <- t(x_c) %*% x_c
40    txy <- t(x_c) %*% y_c
41    iter <- 0
42    juli <- 1
43    ############### loop begin
44    while( juli > eps ){##eps <- 0.1 stop
45      u_beta <- lam * 1 / (abs(beta_ols) * abs(BB))
46      U_beta <- diag(u_beta)
47      BB_new <- solve( txx + U_beta) %*% txy
48      cha  <-  BB_new  -  BB
```

```
49      juli  <- ( t(cha) %*% cha ) ^ 0.5
50      iter <- iter + 1
51      BB   <- BB_new
52      if(iter > itemax) break;
53    }
54    ############loop end
55    beta0 <- c(y_mean - x_mean %*% BB)
56    index_zero <- which(BB <= thre)
57    BB_threshold <- BB
58    BB_threshold[index_zero] <- 0
59
60    beta_full <- c(beta0, BB_threshold)
61    names(beta_full) <- c("Intercept", x_colname)
62    res <- y - (beta0 + x %*% BB_threshold)
63    BIC_my <- log(sum(res^2)) + log(n) / n * sum(BB_threshold!=0)
64    return(list(beta=beta_full, BIC=BIC_my, iter=iter))
```

```
65    ##################################inclued intercept end
66  }else{
67    ##################################not inclued intercept begin
68    txx <- t(x) %*% x
69    txy <- t(x) %*% y
70    iter <- 0
71    juli <- 1
72    ##########loop begin
73    while( juli > eps ){##eps <- 0.1 stop
74      u_beta <- lam * 1 / (abs(beta_ols) * abs(BB))
75      U_beta <- diag(u_beta)
76      BB_new <- solve( txx + U_beta) %*% txy
77      cha  <-  BB_new  -  BB
78      juli  <- ( t(cha) %*% cha ) ^ 0.5
79      iter <- iter + 1
80      BB   <- BB_new
```

```
81        if(iter > itemax) break;
82      }
83      ########loop end
84      index_zero <- which(BB <= thre)
85      BB_threshold <- BB
86      BB_threshold[index_zero] <- 0
87      beta_full <- c(BB_threshold)
88      names(beta_full) <- x_colname
89      res <- y -  x %*% BB_threshold
90      BIC_my <- log(sum(res^2)) + log(n) / n * sum(BB_threshold!=0)
91      return(list(beta=beta_full, BIC=BIC_my, iter=iter))
92      ##################################not inclued intercept end
93    }
94  }
95
96  adalasso.my(x=x, y=y, intercept=TRUE)
```

```
97  adalasso.my(x=x, y=y, intercept=FALSE)
98
99  library(doParallel)
100 library(foreach)
101
102 cl <- makeCluster(8)
103 registerDoParallel(cl)
104
105 nlam <- 100
106 lam_min <- 0.05
107 lam_max <- 4
108 lam_v   <- seq(from=lam_min, to=lam_max, length=nlam)
109 BIC_lam <- foreach(lam=lam_v, .combine="rbind") %dopar%
110   {adalasso.my(y=y, x=x, lam=lam, intercept=FALSE)$BIC}
111
112 ##plot
```

```
plot(lam_v, BIC_lam)
index_optlam <- min(which(BIC_lam==min(BIC_lam)))
lam_opt <- lam_v[index_optlam]
print(lam_opt)
fit <- adalasso.my(y=y, x=x, lam=lam_opt)
##compared
cbind(fit$beta[-1], beta.ture)
stopCluster(cl)
```

输出的结果

```
> cbind(fit$beta[-1], beta.ture)
             beta.ture
x1 3.01404086       3.0
x2 1.50588042       1.5
x3 0.01028767       0.0
x4 0.00000000       0.0
```

x5	2.00791256	2.0
x6	0.00000000	0.0
x7	0.00000000	0.0
x8	0.00000000	0.0

6.2.2 坐标下降法

上面的算法不适合 $p > n$ 的情况，下面我介绍一种常见的算法坐标下降法 (Coordinate descent algorithm,**Wu2008**)，该算法一个分量一个分量计算。

给定 $\beta_1^{(k)}, \ldots, \beta_{j-1}^{(k)}, \beta_{j+1}^{(k)}, \ldots, \beta_p^{(k)}$，我们求解

$$\beta_j^{(k+1)} = \arg\min_{b_j} \left\{ Q(\beta_1^{(k)}, \ldots, \beta_{j-1}^{(k)}, \beta_j, \beta_{j+1}^{(k)}, \beta_p^{(k)}) + \lambda_n \frac{1}{2|\beta_j^{(k)}||\hat{\beta}_{ols,j}|} \beta_j^2 \right\} \tag{6.19}$$

其中 $Q(\beta)$ 是 $\|y - \boldsymbol{x}\beta\|^2$。算法的具体步骤如下：**大家可以尝试编写一下代码试试)**

- 第一步：初始值 $\beta^{(0)}$
- 第二步：$k \geq 0$, 给定 $\beta^{(k)}$

2.1 对于 $j \in 1, \ldots, p$, 利用 (6.19) 更新 $b_j^{(k+1)}$

2.2 重复上面不步骤直到 $b_j^{(k+1)}$ 收敛，从得到 $\beta^{(k+1)}$

- 第三步：重复第二步直到 $\beta^{(k)}$ 收敛。

6.3 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 7 讲 广义线性模型

内容提要

- 广义指数族
- 广义线性模型
- Newton Rashon

7.1 指数分布族

设 $\boldsymbol{Y}$ 是随机变量，如果其密度函数或者质量函数（Density Function or mass function) 具有如下形式

$$f(y;\theta,\phi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right\} \tag{7.1}$$

其中

- θ 是自然参数（Natural parameter）

- ϕ 是冗余参数（Nuisance or dispersion parameter）
- $a(\cdot), b(\cdot)$ 和 $c(\cdot,\cdot)$ 已知的函数

这称 Y 服从指数分布族（Exponential family of distributions）。

自然参数也是我们关心的参数，或者说需要直接估计的参数。冗余参数不是我们直接感兴趣。类似于极大似然估计正态分布的均值喝方差。均值是直接关系的参数，需要直接估计；而方差不是直接关系的参数，利用均值估计可以得到方差估计。

正态分布、二项分布、Bernoul 分布和 Poisson 分布都属于广义指数族。下面我们简单介绍。

例题 **7.1** 正态分布 $Y \sim N\left(\mu, \sigma^2\right)$，则

$$
\begin{aligned}
f(y;\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left(y^2 - 2y\mu + \mu^2\right) - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right\} \\
&= \exp\left\{\left(y\mu - \frac{\mu^2}{2}\right)/\sigma^2 + \left(-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right)\right\}
\end{aligned}
\tag{7.2}
$$

对比 (7.1),

$$\theta = \mu$$
$$a(\phi) = \sigma^2$$
$$b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$$
$$c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)$$

例题 **7.2** 二项分布 **(Binomial distribution)**

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

$$f(y; \pi) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y} \tag{2.3}$$

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

$$f = \exp\left\{ y \log \pi + (n - y) \log(1 - \pi) + \log \binom{n}{y} \right\} \tag{2.4}$$
$$= \exp\left\{ y \log \frac{\pi}{1 - \pi} + n \log(1 - \pi) + \log \binom{n}{y} \right\}$$

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

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

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

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

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

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

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

Note that (2.7) can be written as

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

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

7.1.1 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

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

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

$$\begin{aligned}
\frac{d}{d\theta}f &= f\left\{\frac{d}{d\theta}\log f\right\} && (2.12)\\
&= f(y;\theta,\phi)\frac{d}{d\theta}\left\{\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right\} && (2.13)\\
&= \frac{f(y;\theta,\phi)\,(y - b'(\theta))}{a(\phi)}
\end{aligned}$$

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

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

so that we obtain

$$\begin{aligned} 0 &= \frac{1}{a^2(\phi)} \int_{-\infty}^{+\infty} (y - b'(\theta))^2 f(y;,\theta,\phi) dy - \frac{b''(\theta)}{a(\phi)} \int_{-\infty}^{+\infty} f(y;,\theta,\phi) dy \\ &= \frac{1}{a^2(\phi)} \mathrm{Var}(Y) - \frac{b''(\theta)}{a(\phi)} \end{aligned}$$

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

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

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

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

$$\mathrm{Var}(\boldsymbol{Y}) = b''(\theta)a(\phi) = \exp\{\theta\} = \lambda.$$

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,

$$E(U) = \frac{E(Y) - b'(\theta)}{a(\phi)} = 0$$

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

$$-\frac{\partial U}{\partial \theta} = \mathrm{Var}(U)$$

$\Longrightarrow$

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

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 $\boldsymbol{\theta}$.

7.2 Generalised linear models (GLMs)

7.2.1 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

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}_{n\times 1}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}_{n\times 1}, \quad \text{and} \quad X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}_{n\times p} \tag{7.3}$$

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

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.,

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

7.2.2 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.

7.3 Estimation procedure in GLMs

7.3.1 General principles of finding MLEs

Suppose we have the log-likelihood function of unknown parameters θ, say $l(\theta)$. We want to find out the maximum likelihood estimates (MLEs) $\hat{\theta}$ of θ, i.e.,

$$\hat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}\in\Omega}\{l(\boldsymbol{\theta})\}$$

which is the solution of estimating equation:

$$\frac{\partial l(\theta)}{\partial \theta} = 0.$$

1. Under certain special circumstances, for example, Normal distribution the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ may have explicit mathematical expressions.
2. In general, however, there is no explicit mathematical solution for the MLE $\hat{\theta}$. Certain numerical optimization methods are needed.
3. Two most commonly used optimization methods are Newton-Raphson algorithm and Fisher scoring algorithm. Both algorithms involve the calculation of the first two derivatives of $l(\theta)$

with respect to θ.

(a) **Newton-Raphson algorithm**

The algorithm iteratively calculates the MLE $\hat{\theta}$ through

$$\theta^{(m)} = \theta^{(m-1)} + \left[-l''\left(\theta^{(m-1)}\right)\right]^{-1}\left[l'\left(\theta^{(m-1)}\right)\right] \tag{4.1}$$

where

$$l'\left(\theta^{(m-1)}\right) = \left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\theta^{(m-1)}}$$

$$l''\left(\theta^{(m-1)}\right) = \left.\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'}\right|_{\theta=\theta^{(m-1)}}$$

are $p \times 1$ and $p \times p$ vector/matrix, p is the dimension of θ, and $m = 1, 2, \cdots$ is the iteration number.

Note 1:

$$l'(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \left(\frac{\partial l(\theta)}{\partial \theta_i}\right)_{1 \le i \le p}$$

is called the **Score function** of $\boldsymbol{\theta}$, and

$$[-l''(\theta)] = -\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'} = \left(-\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j}\right)_{1 \le i,j \le p}$$

is called the **observed information matrix of** θ.

Note 2: The algorithm (4.1) requires an initial value of θ, say $\boldsymbol{\theta}^{(0)}$, to start the iteration process.

Note 3: The algorithm (4.1) iterates until convergence. For example, when the solution satisfies

$$\frac{\left\|\theta^{(m)} - \theta^{(m-1)}\right\|}{\left\|\theta^{(m-1)}\right\|} \leq 10^{-5}$$

where $\|\theta\|$ means the squared norm $\|\theta\| = \sqrt{\theta'\theta}$, then the iterations stop and $\boldsymbol{\theta}^{(m)}$ can be viewed as the MLE $\hat{\boldsymbol{\theta}}$.

(b) **Fisher-scoring algorithm**

The Fisher-scoring algorithm is the same as the Newton Raphson algorithm except that the observed information matrix $[-l''(\theta)]$ in (4.1) is replaced by the Fisher information matrix

$$I(\theta) = E\left[-l''(\theta)\right] = -\int l''(\theta \mid Y) f_Y(Y \mid \theta) dY$$

where the expectation is taken with respect to the random response variable $\boldsymbol{Y}$.

Note 4: The Fisher-scoring algorithm and the Newton-Raphson algorithm usually converge to the same solution. The former, in some circumstances, might be simpler than the latter in terms of analytic form of information matrix. For example, the Fisher information matrix might be diagonal but the observed information matrix might be not.
Note 5: Both algorithms provide the estimated covariance matrix of the MLE $\hat{\theta}$.

7.3.2 MLEs in GLMs

Suppose we have an independent random sample $Y_1, Y_2, \cdots, Y_n$ satisfying the properties of a GLM, that is, $Y_1, Y_2, \cdots, Y_n$ have an exponential family of distributions and there is a link function such that $g(\mu_i) = x_i^T \beta$ where $\mu_i = E(Y_i)$. How to find out the MLE of β in the GLM?
First, according to the distribution assumption in (2.1), we know the log-likelihood function in the GLM is

$$l = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \frac{(y_i \theta_i - b(\theta_i))}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)$$

Denote $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_p\right)^T$. The score function of $\boldsymbol{\beta}_j$ is

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\left(y_i - b'\left(\theta_i\right)\right)}{a_i(\phi)} \frac{\partial \theta_i}{\partial \beta_j} \tag{4.2}$$
$$= \sum_{i=1}^{n} \frac{\left(y_i - \mu_i\right)}{a_i(\phi)} \frac{\partial \theta_i}{\partial \beta_j}$$

where $\mu_i = \boldsymbol{E}\left(\boldsymbol{Y}_i\right) = \boldsymbol{b}'\left(\boldsymbol{\theta}_i\right)$, see, e.g. (2.9). Using the chain rule of differentiation we have

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$$

Since

$$\frac{\partial \theta_i}{\partial \mu_i} = 1 / \frac{\partial \mu_i}{\partial \theta_i} = \frac{1}{b''\left(\theta_i\right)} = \frac{a_i(\phi)}{b''\left(\theta_i\right) a_i(\phi)} = \frac{a_i(\phi)}{\operatorname{Var}\left(Y_i\right)}$$

and

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_j} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_j} x_{ij}$$

where x_{ij} is the j-th component of the vector x_i, $j = 1, 2, \cdots, p$, we know

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{a_i(\phi)}{\text{Var}\,(Y_i)} \frac{\partial \mu_i}{\partial \eta_j} x_{ij}$$

Therefore, (4.2) becomes

$$\begin{aligned} U_j &= \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\text{Var}\,(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \\ &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{g'(\mu_i)\, V_i} x_{ij} \end{aligned} \tag{4.3}$$

where $\frac{\partial \mu_i}{\partial \eta_i} = 1/\frac{\partial \eta_i}{\partial \mu_i} = 1/g'(\mu_i)$ as $\eta_i = g(\mu_i)$, and $V_i \equiv \text{Var}\,(Y_i)$. Hence the score vector of β is

$$U \equiv U(\beta) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{g'(\mu_i)\, V_i} x_i \tag{4.4}$$

On the other hand, the derivative of U_j in (4.3) with respect to $\boldsymbol{\beta}_k$ is

$$\begin{aligned}\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} = \frac{\partial U_j}{\partial\beta_k} &= \sum_{i=1}^{n}\left(-\frac{\partial\mu_i}{\partial\beta_k}\right)\frac{1}{g'(\mu_i)V_i}x_{ij} \\ &+ \sum_{i=1}^{n}(y_i-\mu_i)\frac{\partial}{\partial\beta_k}\left[\frac{1}{g'(\mu_i)V_i}\right]x_{ij} \\ &(j,k=1,2,\cdots,p).\end{aligned} \tag{4.5}$$

The second term in 4.5 vanishes when taking expectation from both sides due to $\boldsymbol{E}\left(\boldsymbol{Y_i}-\boldsymbol{\mu_i}\right)=0$. Hence

$$E\left(-\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{n}\frac{1}{[g'(\mu_i)]^2 V_i}x_{ij}x_{ik}$$

where

$$\frac{\partial\mu_i}{\partial\beta_k} = \frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_k} = \left[1/\frac{\partial\eta_i}{\partial\mu_i}\right]\left[\frac{\partial\eta_i}{\partial\beta_k}\right] = \frac{1}{g'(\mu_i)}x_{ik},$$

which is the (j,k)-th element of $p\times p$ Fisher information matrix $\boldsymbol{I}(\boldsymbol{\beta})$. It can be written in matrix

form as

$$I(\beta) = E\left(\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right) = \sum_{i=1}^{n} \frac{1}{\left[g'\left(\mu_i\right)\right]^2 V_i} x_i x_i^T \tag{4.6}$$

Furthermore, if we denote $W_i = 1/\left[g'\left(\mu_i\right)\right]^2 V_i$ and

$$\begin{aligned} \boldsymbol{W} &= \operatorname{diag}\left(\boldsymbol{W}_i, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_n\right) \\ &= \begin{pmatrix} \boldsymbol{W}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{W}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{W}_n \end{pmatrix} \end{aligned}$$

Then (4.6) can be expressed as

$$\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{W} \boldsymbol{X} \tag{4.7}$$

where $X = \left(x_1, x_2, \cdots, x_n\right)^T$ is the $n \times p$ design matrix.

Let $\boldsymbol{D} = \operatorname{diag}\left(\boldsymbol{g'}\left(\boldsymbol{\mu}_1\right), \boldsymbol{g'}\left(\boldsymbol{\mu}_2\right), \cdots, \boldsymbol{g'}\left(\boldsymbol{\mu}_n\right)\right)$, the $\boldsymbol{n} \times \boldsymbol{n}$ diagonal matrix with (i,i)-th element

$g'(\mu_i), \quad (i \le i \le n)$. The score function in (4.4) can be written as

$$U = U(\beta) = X^T W D(y - \mu) \tag{4.8}$$

where $y = (y_1, y_2, \cdots, y_n)^T$ and $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^T$

Suppose we already have an estimate $\boldsymbol{\beta}^{(m-1)}$. Based on this we calculate

$$\mu^{(m-1)} = \mu\left(\beta^{(m-1)}\right), \quad W^{(m-1)} = W\left(\beta^{(m-1)}\right)$$

and

$$D^{(m-1)} = D\left(\beta^{(m-1)}\right)$$

then the Fisher-scoring algorithm indicates the next iteration of $\boldsymbol{\beta}$ is

$$\begin{aligned}\beta^{(m)} &= \beta^{m-1} + \left[I\left(\beta^{(m-1)}\right)\right]^{-1}\left[U\left(\beta^{(m-1)}\right)\right] \\ &= \beta^{m-1} + \left[X' W^{(m-1)} X\right]^{-1}\left[X' W^{(m-1)} D^{(m-1)}\left(y - \mu^{(m-1)}\right)\right]\end{aligned}$$

The right hand side can be written as

$$\left[\boldsymbol{X}'\boldsymbol{W}^{(m-1)}\boldsymbol{X}\right]^{-1}\left[\boldsymbol{X}'\boldsymbol{W}^{(m-1)}\boldsymbol{X}\boldsymbol{\beta}^{(m-1)}+\boldsymbol{X}'\boldsymbol{W}^{(m-1)}\boldsymbol{D}^{(m-1)}(\boldsymbol{y}-\right.$$

$$=\left[\boldsymbol{X}'\boldsymbol{W}^{(m-1)}\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{W}^{(m-1)}\left[\boldsymbol{X}\boldsymbol{\beta}^{(m-1)}+\boldsymbol{D}^{(m-1)}\left(\boldsymbol{y}-\boldsymbol{\mu}^{(m-1)}\right)\right]$$

Denote $\boldsymbol{Z}^{(m-1)}=\boldsymbol{X}\boldsymbol{\beta}^{(m-1)}+\boldsymbol{D}^{(m-1)}\left(\boldsymbol{y}-\boldsymbol{\mu}^{(m-1)}\right)$, then $\boldsymbol{\beta}^{(m)}$ can be rewritten as

$$\boldsymbol{\beta}^{(m)}=\left(\boldsymbol{X}'\boldsymbol{W}^{(m-1)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{(m-1)}\boldsymbol{Z}^{(m-1)}, \tag{4.9}$$

where $m=1,2,\cdots$

Note 1. (4.9) implies that, given a solution of the parameter $\boldsymbol{\beta}$, we first calculate the "working weight matrix" $\boldsymbol{W}$ and the "working response vector" Z. Then the solution of $\boldsymbol{\beta}$ can be updated using the weighted least square method.

Note 2. Since the "working weight matrix" W and the "working response vector" Z depend on the previous solution of $\boldsymbol{\beta}$, **4**.9 is called "iteratively weighted least squares" (IWLS) estimation procedure.

Note 3. The key result of this subsection can be summarized as

Theorem: For the GLMs, the MLE of the regression coefficients $\boldsymbol{\beta}$ can be obtained using the IWLS estimation procedure (4.9).

Note 4. The above theorem is based on Fisher-scoring algorithm. If Newton-Raphson algorithm is used, the solution may be more complicated because of the form of the observed information matrix.

Note 5. There is a special case in which the observed information matrix is completely identical to the Fisher information matrix, that is, the link function is canonical.

In fact, when the link is canonical, i.e., $\boldsymbol{\theta}_i = \boldsymbol{\eta}_i$, we know that $\boldsymbol{\mu}_i = \boldsymbol{E}\left(\boldsymbol{\mu}_i\right) = \boldsymbol{b}'\left(\boldsymbol{\theta}_i\right) = \boldsymbol{b}'\left(\boldsymbol{\eta}_i\right)$. On the other hand, since $\boldsymbol{g}\left(\boldsymbol{\mu}_i\right) = \boldsymbol{\eta}_i$ we have $\boldsymbol{\mu}_i = \boldsymbol{g}^{-1}\left(\boldsymbol{\eta}_i\right)$ (inverse function) so that

$$g^{-1}\left(\eta_i\right) = b'(\eta)$$

Therefore, we obtain

$$b''\left(\eta_i\right) = \left[g^{-1}\left(\eta_i\right)\right]' = \frac{1}{g'\left(\mu_i\right)}$$

i.e., $g'(\mu_i)\, b''(\eta_i) = 1$, and then

$$g'(\mu_i)\, V_i = g'(\mu_i)\, b''(\eta_i)\, a(\phi) = a(\phi)$$

which is independent of the parameter $\boldsymbol{\beta}$.

Obviously, we have

$$\frac{\partial}{\partial \beta}\left[\frac{1}{g'(\mu_i)\, V_i}\right] = 0$$

so that the second term (4.5) disappears if the link is canonical. The first term in (4.5) does not depend on the random variable $\boldsymbol{Y}$. Therefore

$$-\frac{\partial^2 l}{\partial \beta \partial \beta'} = E\left(-\frac{\partial^2 l}{\partial \beta \partial \beta'}\right)$$

as the GLMs have the canonical link. In this case, Newton-Raphson algorithm is the same as Fisher-scoring algorithm.

Note 6 . Although the exponential family of distributions may depend on the dispersion parameter ϕ (or $a(\phi)$), the MLE of the regression coefficients $\boldsymbol{\beta}$ do not depend on the nuisance parameter $\boldsymbol{\phi}$, because it is eliminated from the IWLS estimation.

Example 16. The artificial data in the table below are counts y observed at various values of a covariate $\boldsymbol{x}$

Table 6: Data for poisson regression example

$\boldsymbol{y_i}$	2	3	6	7	8	9	10	12	15
$\boldsymbol{x}$	−1	−1	0	0	0	0	1	1	1

We use GLM to fit the data, i.e., to explore the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$. Let $\boldsymbol{Y_i}$ be the $\boldsymbol{i}$-th count of the variable $\boldsymbol{Y}$ and denote $\boldsymbol{E}\left(\boldsymbol{Y_i}\right) = \boldsymbol{\mu_i}$. We build the model via

$$g\left(\mu_i\right) = x_i^T \beta$$

For the Poisson data set, we choose the canonical link, i.e., the log-link

$$\log \mu_i = \beta_0 + \beta_1 x_i = (1, x_i) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = x_i^T \beta$$

where $\boldsymbol{x_i^T} = (1, \boldsymbol{x}_i)\,, \boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$

Next, we want to see the expressions of $\boldsymbol{W}$ and $\boldsymbol{Z}$. Since $\boldsymbol{g}\left(\boldsymbol{\mu_i}\right) = \log \mu_i$, we have $g'\left(\mu_i\right) = 1/\mu_i$.

For Poisson distribution, it is obvious that

$$\boldsymbol{V}_i = \boldsymbol{Var}\left(\boldsymbol{Y}_i\right) = \boldsymbol{E}\left(\boldsymbol{Y}_i\right) = \mu_i$$

so that

$$W_i = \left[\left(g'\left(\mu_i\right)^2 V_i\right)\right]^{-1} = 1/\frac{1}{\mu_i^2}\mu_i = \mu_i = \exp\left\{x_i^T\beta\right\}$$

and

$$Z_i = x_i^T\beta + g'\left(\mu_i\right)\left(y_i - \mu_i\right) = x_i^T\beta + \left(y_i - \mu_i\right)/\mu_i$$

Furthermore, we have

$$\begin{aligned}
\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} &= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}^T \begin{pmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & W_n \end{pmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n W_i & \sum_{i=1}^n W_i x_i \\ \sum_{i=1}^n W_i x_i & \sum_{i=1}^n W_i x_i^2 \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n \exp\left\{x_i^T\beta\right\} & \sum_{i=1}^n x_i \exp\left\{x_i^T\beta\right\} \\ \sum_{i=1}^n x_i \exp\left\{x_i^T\beta\right\} & \sum_{i=1}^n x_i^2 \exp\left\{x_i^T\beta\right\} \end{pmatrix}
\end{aligned}$$

so that

$$\begin{aligned}
(X'WX)^{-1} &= \begin{pmatrix} \sum_{i=1}^n x_i^2 \exp\left\{x_i^T\beta\right\} & -\sum_{i=1}^n x_i \exp\left\{x_i^T\beta\right\} \\ -\sum_{i=1}^n x_i \exp\left\{x_i^T\beta\right\} & \sum_{i=1}^n \exp\left\{x_i^T\beta\right\} \end{pmatrix} \\
&/\left\{\left[\sum_{i=1}^n \exp\left\{x_i^T\beta\right\}\right]\left[\sum_{i=1}^n x_i^2 \exp\left\{x_i^T\beta\right\}\right] - \left[\sum_{i=1}^n x_i \exp\left\{x_i^T\beta\right\}\right]^2\right\}
\end{aligned}$$

On the other hand,

$$
\begin{aligned}
(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Z}) &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \boldsymbol{x}_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} \boldsymbol{W}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{W}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{W}_n \end{pmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n \boldsymbol{W}_i \boldsymbol{Z}_i \\ \sum_{i=1}^n \boldsymbol{W}_i \boldsymbol{x}_i \boldsymbol{Z}_i \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n \mu_i \left[x_i^T \beta + (Y_i - \mu_i)/\mu_i \right] \\ \sum_{i=1}^n \mu_i \left[x_i^T \beta + (Y_i - \mu_i)/\mu_i \right] x_i \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n \left[\exp\left\{ x_i^T \beta \right\} x_i^T \beta + \left(Y_i - \exp\left\{ x_i^T \beta \right\} \right) \right] \\ \sum_{i=1}^n \left[x_i \exp\left\{ x_i^T \beta \right\} x_i^T \beta + \left(Y_i - \exp\left\{ x_i^T \beta \right\} \right) x_i \right] \end{pmatrix}
\end{aligned}
$$

If we plot $\log y$ against $\boldsymbol{X}$, we can see that a reasonable initial value for $\beta = (\beta_0, \beta_1)^T$ is,

$$\beta_0 = 2, \quad \beta_1 = 1$$

so that

$$X'WX = \begin{pmatrix} 95.2494 & 54.8200 \\ 54.8200 & 65.6931 \end{pmatrix}$$

and

$$X'WZ = \begin{pmatrix} 222.0694 \\ 152.5132 \end{pmatrix}$$

and then

$$\begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix} = \begin{pmatrix} 95.2494 & 54.8200 \\ 54.8200 & 65.6931 \end{pmatrix}^{-1} \begin{pmatrix} 222.0694 \\ 152.5132 \end{pmatrix} = \begin{pmatrix} 1.9150 \\ 0.7235 \end{pmatrix}$$

The iterative process is continued until it converges. The results are shown in the table below Therefore, the MLE of β is $\beta_0 = 1.8892$ and $\beta_1 = 0.6697$,

Table 7: The results of iterative process

$\boldsymbol{m}$	0	1	2	3	4
$\boldsymbol{\beta}_0^m$	2	1.9150	1.8902	1.8892	1.8892
$\boldsymbol{\beta}_1^m$	1	0.7235	0.6716	0.6697	0.6697

and the inverse of the information matrix $\boldsymbol{I} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ is

$$I^{-1} = \begin{pmatrix} 0.02010 & -0.01419 \\ -0.01419 & 0.03192 \end{pmatrix}$$

which actually is the covariance of $\boldsymbol{\beta}$. It is useful for the confidence interval estimates and also hypothesis test. We will see this later.

Note: A R-code is attached below

$y < -c(2, 3, 6, 7, 8, 9, 10, 12, 15)$

$x < -c(-1, -1, 0, 0, 0, 0, 1, 1, 1)$

$X < -$ cbind(rep(1, 9), x)

beta_o <- c (2, 1)

for (i in 1:100)

{

beta $< -$ beta_0

eta $< -X$%∗ beta

$mu < -$ exp(eta)

```
W <- diag(as.vector(mu))
Z <- X%*% beta + ((y - mu) * mu^(-1))
XWX <- t(X)%*% W%*%X
XWZ <- t(X)%*% W%*%Z
Cov <- solve (XWX)
beta <- Cov %*% XWZ
beta
Cov
XWX
XWZ
}
```

7.4 Statistical Inferences

7.4.1 Goodness-of-fit

1. A good statistical model should be parsimonious in the sense that the number of parameters is small and be accurate in the sense that the fitted values are close to the observations of responses.
2. Given $\boldsymbol{n}$ observations, we fit models containing up to $\boldsymbol{n}$ parameters. A model with the maximum number of parameters is called a **saturated/full model.** For GLMs, saturated models mean that $\dim(\beta) = n$, i.e., $p = n$, the number of regression coefficients is identical to that of observation. In this case, the parameter vector $\boldsymbol{\beta}$ is denoted by $\boldsymbol{\beta}_{full}$ and its MLE is denoted by $\hat{\boldsymbol{\beta}}_{full}$. The log-likelihood of a saturated GLM is denoted as $l\left(\boldsymbol{\beta}_{\boldsymbol{fuul}}\right)$.
3. Another extreme case is the null model that has only one parameter, representing a common constant is used whatever the observation is. In other words, it consigns all the variation in the $\boldsymbol{Y}_s$ to the systematic components.
4. In practice, the null model is too simple and the full model is uninformative as it does not summarize data at all. However, the full model provides a reference for measuring the discrepancy

for an intermediate model with $\boldsymbol{p}$ parameters ($1 \leq \boldsymbol{p} \leq \boldsymbol{n}$).

5. Let $\boldsymbol{l}(\boldsymbol{\beta})$ be the log-likelihood of the intermediate model. Based the idea of likelihood ratio test. The goodness of fit for the intermediate model can be measured using

$$\log \boldsymbol{R} = l\left(\hat{\boldsymbol{\beta}}_{\text{full}}\right) - l(\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}}_{full}$ and $\boldsymbol{l}(\hat{\boldsymbol{\beta}})$ are the MLE of $\boldsymbol{\beta}$ based on the full model and the intermediate model,respectively.

6. Wilk's Theorem guarantees that asymptotic distributions of

$$D = 2 \log R = 2\left[l\left(\hat{\beta}_{full}\right) - l(\hat{\beta})\right] \tag{5.1}$$

is χ^2_{n-p}, the Chi-square distribution with $\boldsymbol{n} - \boldsymbol{p}$ degrees of freedom, under certain conditions. We call $\boldsymbol{D}$ in (5.1) as **Deviance/Residual deviance.**

7. Obviously, the smaller the deviance, the better the intermediate model. For GLMs, assume $\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ are the MLE of the natural parameters $\boldsymbol{\theta}_{\boldsymbol{i}}$ based on the full model and the intermediate model, respectively.

The deviance is

$$\begin{aligned}
\boldsymbol{D} &= 2\left[l\left(\hat{\boldsymbol{\beta}}_{full}\right) - l(\hat{\boldsymbol{\beta}})\right] \\
&= 2\left[\sum_{i=1}^{n} \frac{\left(y_i\tilde{\theta}_i - b\left(\tilde{\theta}_i\right)\right)}{a_i(\phi)} + \sum_{i=1}^{n} c\left(y_i, \phi\right)\right] \\
&- 2\left[\sum_{i=1}^{n} \frac{\left(y_i\hat{\theta}_i - b\left(\hat{\theta}_i\right)\right)}{a_i(\phi)} + \sum_{i=1}^{n} c\left(y_i, \phi\right)\right] \\
&= 2\sum_{i=1}^{n} \frac{\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - \left(b\left(\tilde{\theta}_i\right) - b\left(\hat{\theta}_i\right)\right)\right]}{a_i(\phi)}
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}_i$ are the MLEs of $\boldsymbol{\theta}_i$ in the full model and the intermediate model. If we assume $\boldsymbol{a_i}(\boldsymbol{\phi}) = \boldsymbol{\phi}/\boldsymbol{W_i}$, then the deviance can be written as

$$D = \sum_{i=1}^{n} \frac{2W_i\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - \left(b\left(\tilde{\theta}_i\right) - b\left(\hat{\theta}_i\right)\right)\right]}{\phi} \tag{5.2}$$

where if the link is canonical, then $\tilde{\boldsymbol{\theta}}_i = \tilde{\boldsymbol{\mu}}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{full}$ and $\hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\mu}} = \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}$, respectively. And 5.2 is also called scaled deviance.

Example 17: Deviance for Binomial distribution. Suppose we have a sample $Y_1, Y_2, \cdots, Y_n \sim \mathcal{B}\left(n_i, \pi_i\right)$.

The log-likelihood function is

$$
\begin{aligned}
l(\beta) &= \log \prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \\
&= \sum_{i=1}^{n} \left[y_i \log \pi_i + (n_i - y_i) \log (1-\pi_i) + \log \binom{n_i}{y_i} \right] \\
&= \sum_{i=1}^{n} \left[y_i \log \frac{\pi_i}{1-\pi_i} + \boldsymbol{n}_i \log (1-\pi_i) + \log \binom{n_i}{y_i} \right]
\end{aligned}
$$

1. For a saturated model, the π_i 's are all different so that $\beta = (\pi_1, \pi_2, \cdots, \pi_n)'$ It can be shown that the MLE for the saturated are $\tilde{\pi}_i = \boldsymbol{y}_i / \boldsymbol{n}_i$ Therefore, we have

$$
\begin{aligned}
l\left(\tilde{\beta}_{\text{full}}\right) &= \sum_{i=1}^{n} \left[y_i \log \frac{\tilde{\pi}_i}{1-\tilde{\pi}_i} + n_i \log (1-\tilde{\pi}_i) + \log \binom{n_i}{y_i} \right] \\
&= \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{n_i - y_i} - n_i \log \frac{n_i}{n_i - y_i} + \log \binom{n_i}{y_i} \right]
\end{aligned}
$$

2. For an intermediate model ($1 < p < n$), let $\hat{\pi}_i$ denote the MLE of π_i and then $\hat{y}_i = n_i \hat{\pi}_i$ is the fitted value of $\boldsymbol{y}_i$. The maximized log-likelihood is

$$
\begin{aligned}
l(\hat{\beta}) &= \sum_{i=1}^{n}\left[y_i \log\frac{\hat{\pi}_i}{1-\hat{\pi}_i} + n_i \log\left(1-\hat{\pi}_i\right) + \log\begin{pmatrix} n_i \\ y_i \end{pmatrix}\right] \\
&= \sum_{i=1}^{n}\left[y_i \log\frac{\hat{y}_i}{n_i-\hat{y}_i} - n_i \log\frac{n_i}{n_i-\hat{y}_i} + \log\begin{pmatrix} n_i \\ y_i \end{pmatrix}\right]
\end{aligned}
$$

Therefore the deviance is

$$
\begin{aligned}
D &= 2\left[l\left(\hat{\beta}_{full}\right) - l(\hat{\beta})\right] \\
&= 2\sum_{i=1}^{n}\left[y_i \log\frac{y_i\left(n_i-\hat{y}_i\right)}{\hat{y}_i\left(n_i-y_i\right)} - n_i \log\frac{n_i-\hat{y}_i}{n_i-y_i}\right] \\
&= 2\sum_{i=1}^{n}\left[y_i \log\frac{y_i}{\hat{y}_i} - \left(n_i-y_i\right)\log\frac{n_i-y_i}{n_i-\hat{y}_i}\right]
\end{aligned}
$$

Example 18: Deviance for a Normal linear model. Suppose we have an independent sample $Y_1, Y_2, \cdots, Y_n$ with $Y_i \sim N\left(\mu_i, \sigma^2\right), (i = 1, 2, \cdots, n)$ where $\mu_i = E\left(Y_i\right) = x_i^T\beta$. Obviously, the log-likelihood function is

$$
l(\beta) = -\frac{1}{2\sigma^2}\left(y_i-\mu_i\right)^2 - \frac{1}{2}n\log\left(2\pi\sigma^2\right)
$$

where σ^2 is known.

1. For the saturated model, i.e., all $\boldsymbol{\mu}_i$'s are different so that $\boldsymbol{\beta} = (\mu_1, \mu_2, \cdots, \mu_n)'$. It can be shown that in this case the MLEs of μ_i are
$$\tilde{\mu}_i = y_i, (i = 1, 2, \cdots, n)$$

Hence the maximized log-likelihood is
$$l(\tilde{\beta}) = -\frac{1}{2} n \log\left(2\pi\sigma^2\right)$$

2. For the intermediate model ($\boldsymbol{p} < \boldsymbol{n}$), it is well-known that the MLEs of $\boldsymbol{\beta}$ are
$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

and the associated maximized log-likelihood is
$$l(\hat{\boldsymbol{\beta}}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - x_i^T \hat{\boldsymbol{\beta}}\right)^2 - \frac{1}{2} n \log\left(2\pi\sigma^2\right)$$

Therefore, the deviance for a Normal linear model is

$$
\begin{aligned}
D &= 2[l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})] \\
&= 2\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - x_i^T\hat{\boldsymbol{\beta}}\right)^2 \\
&= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}})/\sigma^2 \\
&= \boldsymbol{Y}^T\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right)\boldsymbol{Y}/\sigma^2 \\
&= \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}/\sigma^2
\end{aligned}
\tag{5.3}
$$

where $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}$ is a projection matrix (i.e., symmetric and idempotent).

Note 1. Let $e = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ be the residual, then (5.3) implies the deviance for Normal linear model is the residual sum of squares, divided by the variance $\boldsymbol{\sigma}$.

Note 2. Since $Y \sim N_n\left(X\beta, \sigma^2 I_n\right)$, we know that

$$
e = (I_n - H)\,Y \sim N_n\left(0, \sigma^2\,(I_n - H)\right)
$$

as $(\boldsymbol{I_n} - \boldsymbol{H})\, \boldsymbol{X} = 0$. Then

$$\boldsymbol{D} = \boldsymbol{Y}' \left(\boldsymbol{I}_n - \boldsymbol{H}\right) \boldsymbol{Y}/\sigma^2 = e'e/\sigma^2 \sim \chi^2_{n-p}$$

(exactly not asymptotically).

Note 3. The dispersion parameter σ^2 (i.e., the variance) is assumed to be known. If σ^2 is unknown, (5.3) cannot be directly used to measure the goodness of fit. An estimate $\hat{\sigma}^2$ can be used to replace σ^2 in this case.

Example 18: Deviance for a Poisson model. Suppose we have an independent sample $Y_1, Y_2, \cdots, Y_n$ with $Y_i \sim \mathcal{P}(\lambda_i)$, i.e., the Poisson distribution, then the log-likelihood function is

$$l(\beta) = \sum_{i=1}^{n} y_i \log \lambda_i - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \log y_i!$$

1. For the saturated model, i.e., all the λ_i 's are different so that $\boldsymbol{\beta} = (\lambda_1, \lambda_2, \cdots, \lambda_n)'$. It can be show that the MLEs of λ_i are $\lambda_i = y_i$ and then

$$l(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \boldsymbol{y}_i \log \boldsymbol{y}_i - \sum_{i=1}^{n} \boldsymbol{y}_i - \sum_{i=1}^{n} \log \boldsymbol{y}_i!$$

2. For the intermediate model ($p < n$), the MLEs $\hat{\lambda}_i$ can be calculated via $\hat{\lambda}_i = g^{-1}(\hat{\eta}_i) = g^{-1}\left(x_i^T \hat{\beta}\right)$ where $\hat{\beta}$ are the MLEs of β and $\boldsymbol{g}^{-1}(\cdot)$ is the inverse of the link function $\boldsymbol{g}(\cdot)$. Hence the maximized log-likelihood is

$$l(\hat{\beta}) = \sum_{i=1}^{n} y_i \log \hat{y}_i - \sum_{i=1}^{n} \hat{y}_i - \sum_{i=1}^{n} \log y_i!$$

where $\hat{\boldsymbol{y}}_i = \hat{\boldsymbol{\lambda}}_i$ is the fitted value of $\boldsymbol{y}_i$.

Therefore, the deviance is

$$\begin{aligned} D &= 2[l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})] \\ &= 2\left[\sum_{i=1}^{n} \boldsymbol{y}_i \log \frac{\boldsymbol{y}_i}{\hat{\boldsymbol{y}}_i} - \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\right)\right] \end{aligned}$$

Let us look at Example 16. We already obtain the MLEs are $\hat{\boldsymbol{\beta}}_0 = 1.8892$ and $\hat{\boldsymbol{\beta}}_1 = 0.6697$ so that

$$\log \hat{y}_i = \log \hat{\mu}_i = \log \hat{\lambda}_i = 1.8892 + 0.6697 x_i$$

Table 8: Results from the Poisson Regression

x	y_i	$\hat{y}_i$	$y_i \log \frac{y_i}{\hat{y}_i}$	$y_i - \hat{y}_i$
−1	2	3.385448	−2.1053568	−1.3854485
−1	3	3.385448	−0.7252446	−0.3854485
0	6	6.614552	−1.1701503	−0.6145515
0	7	6.614552	0.7929341	0.3854485
0	8	6.614552	3.0427127	1.3854485
0	9	6.614552	5.5431465	2.3854485
1	10	12.923632	−5.1294501	−2.9236323
1	12	12.923632	−1.7796228	−0.9236323
1	15	12.923632	4.4697781	2.0763677
Total	72	72	1.469	0.000

so that $D = 2 \times 1.469 = 2.9387$, which is smaller than $\chi^2_{7,0.05} = 14.07$ (the lower 5% tail of the distribution χ^2_7). It indicates that the model fit the data very well. Or p-value = 0.8906 (i.e., $\Pr\left\{\chi^2_7 \geq 2.9387\right\} = 0.8906$). Other important measure of goodness-of-fit is the Generalized

Pearson χ^2 statistic, defined by

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

where $\hat{\mu}_i$ and $\boldsymbol{V}(\hat{\mu}_i)$ are the estimated mean and variance. For example, for Poisson distribution $Y_i \sim \mathcal{P}(\lambda_i)$ we have

$$\hat{\mu}_i = \hat{\lambda}_i = g^{-1}\left(x_i^T \beta\right)$$

and

$$V(\hat{\mu}_i) = \hat{\mu}_i = \hat{\lambda}_i$$

1. It can also be shown that in GLMs $\chi^2 \sim \chi^2_{n-p}$, where p is the dimension of the regression coefficients.

2. We can show that

(a) $\mathcal{B}(\boldsymbol{n}_i, \pi_i) : \chi^2 = \sum_{i=1}^{n} \frac{(y_i - n_i \hat{\pi}_i)^2}{n_i \hat{\pi}_i (1-\hat{\pi}_i)}$

(b) $\mathcal{B}(\pi_i) : \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i (1-\hat{\pi}_i)}$

(c) $\mathcal{P}(\lambda_i) : \chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i}$

(d) $N\left(\mu_i, \sigma^2\right) : \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}, \quad \left(\hat{\mu}_i = x_i^T \hat{\beta}\right)$

Interpretation of deviance: Since $D \sim \chi^2_{n-p}$, it is concluded that if $D < \chi^2_{n-p;\alpha}$ (the critical value of χ^2 with degree of freedom $n - p$ at level α) then the intermediate model is not significantly different from the full model in terms of actuary. In other words, the intermediate model can be viewed as good as the full model in terms of accuracy. In this case, the intermediate model is acceptable. Otherwise, the model is not good enough.

Other residuals:

1. Pearson residual: Pearson residual is defined as

$$\boldsymbol{r}_{pi} = \frac{\boldsymbol{y}_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

and then

$$\chi^2 = \sum_{i=1}^{n} r_{pi}^2$$

2. Deviance residual: Recall the definition of deviance

$$
\begin{aligned}
D &= 2\left[l\left(\hat{\boldsymbol{\beta}}_{\text{full}}\right) - l(\hat{\boldsymbol{\beta}})\right] \\
&= 2\sum_{i=1}^{n} \frac{\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - \left(b\left(\tilde{\theta}_i\right) - b\left(\hat{\theta}_i\right)\right)\right]}{a_i(\phi)}
\end{aligned}
$$

Let $d_i = 2\dfrac{\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - \left(b\left(\tilde{\theta}_i\right) - b\left(\hat{\theta}_i\right)\right)\right]}{a_i(\phi)}$, then we have $d_i > 0$. Define

$$
r_{Di} = \text{sign}\left(y_i - \mu_i\right)\sqrt{d_i}
$$

where $\text{sign}(\cdot)$ is the sign function, i.e.,

$$
\text{sign}(t) = \begin{cases} 1, & \text{if } t > 0 \\ -1, & \text{otherwise} \end{cases}
$$

Obviously we have

$$
\boldsymbol{D} = \sum_{i=1}^{n} r_{Di}^2
$$

3. General interpretation of residuals: If the residuals are small in the sense of absolute value, then the model fits data well. Otherwise, the fitting is not good enough.

7.4.2 Hypothesis testing

For a GLM, we may be concerned with the significance of some regression coefficients. Consider the following typical hypothesis

$$\boldsymbol{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \text{ against } \boldsymbol{H}_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 \tag{5.4}$$

where

$$\boldsymbol{\beta}_0 = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_q\right)'$$
$$\boldsymbol{\beta}_1 = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_q, \boldsymbol{\beta}_{q+1}, \cdots, \boldsymbol{\beta}_p\right)'$$

($q < p < n$). We would like to know if the model under the null hypothesis $\boldsymbol{H}_0$ is as good as the one under the alternative hypothesis $\boldsymbol{H}_1$.

1. The general criterion based on likelihood ratio test can be used to test the hypothesis in (5.4).
2. For GLMs, the deviance can be used to test (5.4) too, and it is actually much better than the likelihood ratio test. Assume D_0 and D_1 are the deviances based on H_0 and H_1, respectively, i.e.,

$$D_0 = 2\left[l\left(\hat{\beta}_{\text{full}}\right) - l\left(\hat{\beta}_0\right)\right]$$
$$D_1 = 2\left[l\left(\hat{\beta}_{\text{full}}\right) - l\left(\hat{\beta}_1\right)\right]$$

where $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ are the MLE of $\boldsymbol{\beta}$ under $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$, respectively. Then their difference is

$$\Delta D = D_0 - D_1 = 2\left[l\left(\hat{\beta}_1\right) - l\left(\hat{\beta}_0\right)\right] \tag{5.5}$$

If both models fit the data well then we have

$$D_0 \sim \chi^2_{n-q} \text{ and } D_1 \sim \chi^2_{n-p}$$

so that

$$\Delta D \sim \chi^2_{p-q}$$

If $\Delta D > \chi^2_{p-q;\alpha'}$, we reject the null hypothesis H_0 in favor of H_1 (even if it may not fit the data well). Otherwise, we do not have evidence to reject the null hypothesis $\boldsymbol{H_0}$.

In some cases, distributions in GLMs may involve unknown nuisance parameters and the deviance difference in 5.5) may not be useful for hypothesis test. However, it can be modified to eliminate the nuisance parameters in terms of F-statistics.

Example 19: Hypothesis testing for a Normal linear model. For the Normal linear model $E(Y_i) = x_i^T \beta$, it was shown that the deviance is

$$D = \sum_{i=1}^{n} \frac{\left(y_i - x_i^T \hat{\beta}\right)^2}{\sigma^2}$$

so that

$$\Delta D = \sum_{i=1}^{n} \frac{\left[\left(y_i - x_i^T \hat{\beta}_0\right)^2 - \left(y_i - x_i^T \hat{\beta}_1\right)^2\right]}{\sigma^2}$$

where σ^2 is the variance, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the MLEs of $\boldsymbol{\beta}$ under $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$, respectively.
Although $\Delta \boldsymbol{D}$ in theory is asymptotically distributed as χ^2_{p-q}, it can be calculated only when σ^2 in

know. When it is unknown, it can be eliminated using certain technique, for example, by

$$
\begin{aligned}
F &= \frac{D_0 - D_1}{p-q} \Big/ \frac{D_1}{n-p} \\
&= \frac{\sum_{i=1}^{n} \left[\left(y_i - x_i' \hat{\beta}_0\right)^2 - \left(y_i - x_i' \hat{\beta}_1\right)^2 \right] / (p-q)}{\sum_{i=1}^{n} \left[\left(y_i - x_i' \hat{\beta}_1\right)^2 \right] / (n-p)} \\
&\sim F_{p-q,n-p}
\end{aligned}
\tag{5.6}
$$

provided $\boldsymbol{H}_0$ is true. Therefore, (5.6) can be used to carry out the hypothesis testing about $\boldsymbol{\beta}$, regardless of whether or not the variance σ^2 is known.

第 8 讲 多分类 Logit 模型

内容提要

- 多项分布
- 对数线性模型
- 多分类
- 多分类 Logit 模型
- Iris 数据

8.1 对数线性模型

8.1.1 多项分布

假设在 n 次独立同分布实验，每次实验结果是 J 中可能类别的某一类。假设第 i 次实验的结果为类别 j，则 $y_{ij}=1$，否则 $y_{ij}=0$。此时 $\boldsymbol{y}_i=(y_{i1},\ldots,y_{iJ})$ 表示一个多项式分布实验，且 $\sum_{j=1}^{J} y_{ij}=1$。由于这个限制，所以某一类的值可以有其他类的值线性表示，假设 y_{iJ} 可以

有其他 $y_{ij}(j=1,\ldots,J-1)$ 线性表示，即 $y_{iJ}=1-\sum_{j=1}^{J-1}y_{ij}$，或句话说 y_{iJ} 是冗余的。n 次试验中，第 j 类别出现的次数为 $n_j=\sum_{i=1}^{n}y_{ij}$，那么 $(n_1,\ldots,n_J)$ 服从多项分布 (Multinomial distribution)，其概率质量函数为

$$p(n_1,\ldots,n_J)=\frac{n!}{n_1!\ldots n_J!}\pi_1^{n_1}\ldots\pi_J^{n_J} \tag{8.1}$$

其中 $P(Y_{ij}=1)=\pi_j$，$\sum_{j=1}^{J}n_i=n$。其分布的参数为 $(n,\pi_1,;\pi_J)$。当 $J=2$ 是多项分布退化为二项分布。

下面，我们讨论泊松分布与多项分布的额关系。假设 $Y_i\sim P(\mu_j)$，并且 $Y_1,\ldots,Y_J$ 相互

独立，则

$$
\begin{aligned}
&P\Big(Y_1 = n_1, \ldots, Y_J = n_J \Big| \sum_j^n Y_j = n\Big) \\
&= \frac{P(Y_1 = n_1, \ldots, Y_J = n_J)}{P(\sum_{j=1}^J Y_j = n)} \\
&= \frac{\prod_{j=1}^J \Big(\exp(-\mu_j)\pi_j^{n_j}/n_j\Big)}{\exp(-\sum_{j=1}^J \mu_i)\Big(\sum_{j=1}^J \pi_j\Big)/n!} \\
&= \frac{n!}{\prod_{j=1}^J n_j} \prod_{j=1}^J \Big(\frac{\mu_j}{\sum_{j=1}^J \mu_j}\Big)^{n_j} \\
&= \frac{n!}{\prod_{j=1}^J n_j} \prod_{j=1}^J \pi_j{}^{n_j}
\end{aligned}
$$

最后一项恰好是参数为 $(n, \pi_1, \ldots, \pi_J)$ 的多项分布。这也暗示实际中多分类的问题中，基于泊松分布和多项分布模型的参数是一样的。两者某种程度可以相互转化。

8.2 多分类 Logit 模型

8.2.1 模型表达

令 Y 的类别个数为 J，当有 n 个独立的观测时，这 J 种类型结果的个数的概率分布就是多项分布。这也是多分类 Logit 模型也称为多项式 Logit 模型的原因。

假设为 $\pi_1(\boldsymbol{x}) = P(Y = j|\boldsymbol{X} = \boldsymbol{x})$ 表示给定自变量的取值 $\boldsymbol{x}$ 下 Y 是 j 的概率。给定 $\boldsymbol{x}$，我们将 Y 在 J 的概率为 $\{\pi_1(\boldsymbol{x}), \ldots, \pi_J(\boldsymbol{x})\}$。通常采用对数发生比处理，即将每一个类别座位基线（Baseline），其他的类别与其组成发生比进行建模，也称为基线-类别 logit 模型，其表示为：

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}_j, j = 1, \ldots, J-1. \tag{8.2}$$

这里我们需要注意，这个模型有一个假设，$J-1$ 的发生比模型中自变量是相同的。每个模型有不同的参数. 这些效应根据基线配对的类别而变化.

模型（8.2）其实可以决定任意两个类别的发生比，假设对于任意两个类别 k 和 l，

$$\begin{aligned}\log\frac{\pi_k(\boldsymbol{x})}{\pi_l(\boldsymbol{x})} &= \log\frac{\pi_k(\boldsymbol{x})/\pi_J(\boldsymbol{x})}{\pi_l(\boldsymbol{x}/\pi_J(\boldsymbol{x})}\\ &= \log\frac{\pi_k(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} - \log\frac{\pi_l(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\\ &= \alpha_k + \boldsymbol{x}^\top\boldsymbol{\beta}_k - \alpha_l - \boldsymbol{x}^\top\boldsymbol{\beta}_l\\ &= (\alpha_k - \alpha_l) + \boldsymbol{x}^\top(\boldsymbol{\beta}_k - \boldsymbol{\beta}_l)\end{aligned}$$

模型（8.2）可以得到所有分类的条件概率，因为 $\sum_{j=1}^{J}\pi_k(\boldsymbol{x}) = 1$。具体如下：

$$\begin{aligned}\pi_1(\boldsymbol{x}) &= \pi_J(\boldsymbol{x})\exp(\alpha_1 + \boldsymbol{x}^\top\boldsymbol{\beta}_1)\\ \pi_1(\boldsymbol{x}) &= \pi_J(\boldsymbol{x})\exp(\alpha_2 + \boldsymbol{x}^\top\boldsymbol{\beta}_2)\\ &\cdots\\ \pi_{J-1}(\boldsymbol{x}) &= \pi_J(\boldsymbol{x})\exp(\alpha_{J-1} + \boldsymbol{x}^\top\boldsymbol{\beta}_{J-1})\end{aligned}$$

进而得到

$$\pi_J\left(1+(\boldsymbol{x})\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{x}^\top\boldsymbol{\beta}_j)\right)=1$$

所以，我们可以得到

$$\begin{aligned}\pi_j(\boldsymbol{x})&=\frac{\exp(\alpha_j+\boldsymbol{x}^\top\boldsymbol{\beta}_j)}{\left(1+\sum_{j=1}^{j}\exp(\alpha_j+\boldsymbol{x}^\top\boldsymbol{\beta}_j)\right)},\ j=1,\ldots,J-1\\ \pi_J(\boldsymbol{x})&=\frac{1}{\left(1+\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{x}^\top\boldsymbol{\beta}_j)\right)}\end{aligned}\tag{8.3}$$

8.2.2 估计方法

与广义线性模型类似，我们经常使用极大似然估计。设 $\boldsymbol{y}_i=(y_{i1},y_{i2},y_{iJ})^\top$ 表示 i 个观测的多项分布实验，当实验结果落在 j 类别时 $y_{ij}=1$，否则 $y_{ij}=0$。假设有 n 个独立观测

值，则

$$\begin{aligned}
\log \prod_{i=1}^{n}\left[\prod_{i=1}^{n} \pi_j(\boldsymbol{x}_i)\right] &= \sum_{i=1}^{n}\sum_{j=1}^{J}\left[y_{ij}\log \pi_j(\boldsymbol{x}_i)\right] \\
&= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}\log \pi_j(\boldsymbol{x}_i) + \left(1-\sum_{j=1}^{J-1} y_{ij}\right)\log \pi_J(\boldsymbol{x}_i)\right\} \\
&= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}\frac{\pi_j(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)} + \log \pi_J(\boldsymbol{x}_i)\right\} \\
&= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}(\alpha_j + \boldsymbol{x}_i^{\top}\boldsymbol{\beta}_j) - \log\left(1+\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_j)\right)\right\} \qquad (8.4)
\end{aligned}$$

第三个等式成立因为（8.3)。我们可以使用 Newton-Rashon 的方法求解。

8.2.3 Iris 数据分析

Iris（鸢尾花）数据是经典的数据，该数据样本量为 150。因变量是 Species（记为 Y），其有三个类别分别是 Setosa、Versicolour、Virginica（分别记为 1，2，3，此时 $J=3$）。自变量

有 Sepal.Length Sepal.Width Petal.Length Petal.Width(分别记为)。目的是利用四个属性预测鸢尾花卉的种类。从代码结果中，我们可以得到

$$\log \frac{\hat{\pi}_1}{\hat{\pi}_J} = 35.5 + 9.5x_{1i} + 12.3x_{2i} - 23.0x_{3i} - 33.8x_{4i}$$

$$\log \frac{\hat{\pi}_2}{\hat{\pi}_J} = 42.6 + 2.5x_{1i} + 6.7x_{2i} - 9.4x_{3i} - 18.3x_{4i}$$

- 我们是使用 refLevel = "virginica" 设置 virginica 为基线。
- predict(fit) 预测是 $\log \frac{\hat{\pi}_1}{\hat{\pi}_J}$ 和 $\log \frac{\hat{\pi}_2}{\hat{\pi}_J}$。我们利用（8.3）可以得到 $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3$（即 pi_pre)。
- 对于某一观测值，我们已经得到其分类的概率。通常，认为哪一个概率最大其属于哪一类，即利用 which.max 函数实现预测值 $\hat{Y}$(即 y_pre)。
- 利用 table 函数实现准确率的判断。第 1 类全部全部正确识别，第 2 类有一个观测判为第 2 类，第 2 类也有一个观测判为第 2 类。错判率为 2/150 *100=1.3。

笔记

- 国内外很多书，利用 nnet 中的函数 multinom 估计（8.3)。这是错误的，因为其实现是多项式对数线性模型 (Multinomial Log-linear Models)。估计的参数不是（8.3）的参数，这将导致参数解释错误。吴喜之（2019）阐述 multinom 可以实现连续自变量的多项式

对数线性模型：

$$\log(\mu_j) = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}, j = 2, \ldots, J \tag{8.5}$$

其中限制条件是 $n - \sum_{j=1}^{J-1} \mu_j$。前面我们讨论多项式分布和泊松分布的关系 $\pi_j = \mu_j / \sum_{j=1}^{J} \mu_j$。

- `multinom` 的预测值 fitted.values 是 $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_J$。我们利用 summary 给出两个方法预测值差的绝对值的摘要，最大值为 0.001。这暗示几乎没有差异。我们再次强调 multinom 估计（8.5）参数，预测值是 $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_J$。我利用 multinom 求解 μ_j，没有得到合理的结果。这个笔记需要进一步论述。请大家查阅资料

Iris 数据的代码

```
> rm(list=ls())
> library(datasets)
> data(iris)
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
```

```
> summary(iris)
> levels(iris$Species)
[1] "setosa"     "versicolor" "virginica"
> library(VGAM)
> fit <- VGAM::vglm(Species ~.,data =iris,family =multinomial(refLevel = "virginica"))
> fit
Coefficients:
 (Intercept):1  (Intercept):2 Sepal.Length:1 Sepal.Length:2  Sepal.Width:1
     35.490246      42.637804       9.494898       2.465220      12.300446
 Sepal.Width:2 Petal.Length:1 Petal.Length:2  Petal.Width:1  Petal.Width:2
      6.680887     -22.975454      -9.429385     -33.842647     -18.286137
Degrees of Freedom: 300 Total; 290 Residual
Residual deviance: 11.89855
Log-likelihood: -5.949274
> fit_pre <- predict(fit)
> b <- exp(fit_pre)
```

```
> bb <- apply(b, 1, sum)
> pi3 <- 1/ (1+bb)
> pi_pre <- cbind(b * pi3, pi3)
> head(pi_pre)
     p1           p2           p3
[1,]  1 3.798337e-12 5.866136e-39
[2,]  1 2.573092e-10 1.836862e-35
[3,]  1 8.803156e-11 1.053364e-36
> colnames(pi_pre) <- c("p1", "p2", "p3")
> y_pre <- apply(pi_pre, 1, which.max)
> y <- iris$Species
> table(y, y_pre)
            y_pre
y             1  2  3
  setosa     50  0  0
  versicolor  0 49  1
```

```
 virginica    0  1 49
> library(nnet)
> fit <- nnet::multinom(Species ~., data =iris )
> summary(fit)
Call:
nnet::multinom(formula = Species ~ ., data = iris)

Coefficients:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    18.69037    -5.458424   -8.707401     14.24477   -3.097684
virginica    -23.83628    -7.923634  -15.370769     23.65978   15.135301

Std. Errors:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    34.97116     89.89215    157.0415     60.19170    45.48852
virginica     35.76649     89.91153    157.1196     60.46753    45.93406
```

```
Residual Deviance: 11.89973
AIC: 31.89973
> summary(abs(pi_pre - fit$fitted.values))
       p1                   p2                   p3
 Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
 1st Qu.:0.000e+00   1st Qu.:4.900e-09   1st Qu.:0.000e+00
 Median :4.700e-10   Median :3.200e-07   Median :4.900e-09
 Mean   :3.846e-06   Mean   :4.228e-05   Mean   :3.843e-05
 3rd Qu.:4.475e-08   3rd Qu.:5.365e-06   3rd Qu.:3.583e-06
 Max.   :2.456e-04   Max.   :1.179e-03   Max.   :1.179e-03
```

8.3 次序 Logsit 回归

次序变量是含有多个类别的变量，并且类别之间有一定顺序，如满意度（“很满意”、“满意”、和“不满意”）、企业或个人信用评级（“AAAA”、“AAA”，“AA”级等）和身体状况

(“健康”、“亚健康”和“不健康”)等。对于次序响应变量，我们不能将其当作无序响应变量处理，这不仅会破坏数据的自身的规律，也会出现难以解释或与实际意义严重不符的结论。常用的次序 logit 模型主要包括累计 logit 模型（Cumulative logit model)、相邻分类 logit 模型（Adjacent-categories logit model）和连续比 logit 模型（Continuation-ratio logit）

8.3.1 累计 logit 模型

累计 logit 模型一开始提出主要用于关联表，如 Sell(1964)、Clayton(1971) 等。它主要包括成比例优势累计 logit 模型和不成比例优势累计 logit 模型。设因变量有 J 个类别，其发生概率为 $\pi_1,\ldots,\pi_J$，累计 Logit 模型为:

$$Logit(P(Y \le j|X=x)) = \log \frac{P(Y \le j|X=x)}{1-P(Y \le j|X=x)} = \log \frac{\pi_1(\boldsymbol{x})+\cdots+\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})+\cdots+\pi_J(\boldsymbol{x})}, \quad j=1,\ldots,J-1 \tag{8.6}$$

根据自变量的影响，起可以分为

- 成比例优势累计 logit 模型为:

$$Logit(P(Y \le j|X=x)) = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}, \quad j=1,\ldots,J-1 \tag{8.7}$$

- 不成比例优势累计 logit 模型为：

$$Logit(P(Y \le j|X = x)) = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}_j, \quad j = 1, \ldots, \tag{8.8}$$

模型 (8.7) 之所以成比例，因为它满足：

$$\begin{aligned} & Logit(P(Y \le j|\boldsymbol{X} = \boldsymbol{x}_k)) - Logit(P(Y \le j|\boldsymbol{X} = \boldsymbol{x}_l)) \\ &= \log \frac{P(Y \le j|\boldsymbol{X} = \boldsymbol{x}_k)/P(Y > j|\boldsymbol{X} = \boldsymbol{x}_k)}{P(Y \le j|\boldsymbol{X} = \boldsymbol{x}_l)/P(Y > j|\boldsymbol{X} = \boldsymbol{x}_l)} \\ &= (\boldsymbol{x}_k - \boldsymbol{x}_l)^\top \boldsymbol{\beta} \end{aligned}$$

即因变量 $Y < j$ 在 $\boldsymbol{x}_k$ 处的优势（odds）是 $\boldsymbol{x}_l$ 的 $\exp(\boldsymbol{x}_k - \boldsymbol{x}_l)^\top \boldsymbol{\beta}$ 倍。对数累计优势比与 $\boldsymbol{x}_k$ 和 $\boldsymbol{x}_l$ 的距离成比例。

模型（(8.7)）每一个累计 logit 模型，都有自己的截距项。对于固定的自变量，α_j 随着 j 的增大而增大；自变量对所有的累计 logit 模型效应是相同的，从而保证

$$P(Y \le 1|\boldsymbol{X} = \boldsymbol{x}) \le P(Y \le 2|\boldsymbol{X} = \boldsymbol{x}) \le \ldots P(Y \le J|\boldsymbol{X} = \boldsymbol{x}) = 1 \tag{8.9}$$

模型 (8.8) 对于不同的累计 logit 模型，自变量影响不同，但累计 logit 概率可能会在某些自变量值下相交，这显然不合适，违背累计概率的顺序 (8.9)。如果限制不相交，其模型估计极其复杂。另外，当自变量很多时，模型待估参数个数会很多，导致自由度严

重损失。如含有 9 个自变量和 5 个类别响应变量的不成比例优势累计 logit 模型需要估计 40 个参数，而成比例优势累计 logit 模型只需要估计 13 个。前者比后者自由度多损失 27 个。所以，一般来说，不推荐使用不成比例优势累计 logit 模型，除非成比例优势累计 logit 模型拟合不佳，并且结论严重错误。

R 包 VGAM 中 vglm 函数可以实现累计 logit 模型，具体如下

vglm(formula,family=cumulative(link=logit,parallel=TRUE/FALSE), data=name)

- formula 是模型的公式表达（如文章中的实例 formula=cbind(y1,y2,y3) x1+x2）
- family=cummulative 表示累计模型；link=logit 表示连接函数是 logit 函数；
- parallel 表示自变量的效应是否相同，parallel=TURE 表示相同（模型 (8.7)，parallel=FALSE 表示不同（模型 (8.8)）
- data=name 表示数据名是 name。

除了累计 logit 模型，累计模型中还有累计 probit 模型和累计 log-log 模型等。相对应的 R 代码是将 vglm 函数中的 family=cumulative(link=logit) 中的 logit 相应改为 probit、cloglog 即可。

8.3.2 相邻分类 logit 模型

相邻分类 logit 模型有两种形式：

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})}, 1, \ldots, J-1 \tag{8.10}$$

和

$$\log \frac{\pi_{j+1}(\boldsymbol{x})}{\pi_j(\boldsymbol{x})}, 1, \ldots, J-1 \tag{8.11}$$

其中模型 (8.10) 比较常见。相邻分类 logit 模型包括成比例相邻分类 logit 模型和不成比例相邻分类 logit 模型。以模型 (8.10) 为例，成比例相邻分类 logit 模型定义为：

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})} = \alpha_j = \boldsymbol{x}^\top \boldsymbol{\beta}, 1, \ldots, J-1 \tag{8.12}$$

和

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})} \alpha_j = \boldsymbol{x}^\top \boldsymbol{\beta}_j, 1, \ldots, J-1 \tag{8.13}$$

相邻分类 logit 模型与累计 logit 模型类似，前者是利用累计优势，后者是利用局部优势。但是前者不同于后者，它不需要限制条件保证（(8.9)），因为它得到的是相邻类别概率的对数，可以为任意非负值。

相邻分类 logit 模型在 R 中也可以用 vglm 函数实现：

vglm(formula,family= acat(reverse=TRUE/FALSE, parallel=TRUE/FALSE), data=name)

- 当 reverse=TUER，parallel=TURE 表示模型（(8.12)）；reverse=TUER，parallel=FALSE 表示模型（(8.13)
- reverse=FALSE，parallel=TURE 表示模型 (8.11) 的成比例模型；reverse=FALSE，parallel=FALSE 表示模型 (8.11) 的不成比例模型。

8.4 连续比 logit 模型

连续比 logit 模型有两种形式：

- 一种是每个类别相对比自己高的类别对数优势：

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})} \quad j = 1, \ldots, J-1 \tag{8.14}$$

- 另一种是每个类别相对比自己低的类别对数优势：

$$\log \frac{\pi_{j+1}(\boldsymbol{x})}{\pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x})} \quad j = 1, \ldots, J-1 \tag{8.15}$$

与累计 logit 模型类似，模型（8.14）包括两种模型，

- 一种是成比例连续比 logit 模型：

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})} = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta} \quad j = 1, \ldots, J-1 \tag{8.16}$$

- 一种是不成比例连续比 logit 模型：

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})} = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}_j \quad j = 1, \ldots, J-1 \tag{8.17}$$

连续比 logit 模型可用 R 中 vglm 函数实现：

vglm(formula,family= sratio (reverse=TRUE/FALSE, parallel=TRUE/FALSE), data=name)

- 当 reverse=FALSE，parallel=TURE 表示模型（8.16）；reverse=FALSE，parallel=FALSE 表示模型（(8.17)）；
- reverse=TUER，parallel=TURE 表示模型（8.15）的成比例模型；reverse=TUER，parallel=FALSE 表示模型（8.15）的不成比例模型。

8.5 实际数据分析

数据来自 1971 年 Madsen 在 Copenhagen 对住房满意度的调查。针对选定地区在 1960 年到 1968 年建造的出租房内住户，调查他们的住房满意度以及联系其他居民的程度、住房类型（表8.1）。其中 x1 表示联系其他居民，x1=0 代表“少”，x1=1 代表“多”；x2 表示住房类型，x2=1 代表“高层住房”，x2=2 代表“公寓”，x2=3 代表“平房”；y 表示住房满意度，y1 代表“不满意”，y2 代表“比较满意”，y3 代表“非常满意”。为了减少参数个数，文章没有将住房类型用虚拟变量表示，而用得分表示。

表 8.1: 住房满意度

联系其他居民 x1	住房类型 x2	不满意 y1	比较满意 y2	非常满意 y3
少 (0)	高层住房 (1)	65(29.68%)	54(24.65%)	100(45.67%)
少 (0)	公寓 (2)	130(41.00%)	76(23.97%)	111(35.03%)
少 (0)	平房 (3)	67(37.85%)	48(27.12%)	62(35.03%)
多 (1)	高层住房 (1)	34(18.78%)	47(25.97%)	100(55.25%)
多 (1)	公寓 (2)	141(31.47%)	116(25.90%)	191(42.63%)
多 (1)	平房 (3)	130(38.35%)	105(30.97%)	104(30.78%)

由于住房满意度是次序变量，并且有 3 个类别，所以可以建立次序响应变量模型。为了

阐述上面模型的应用，我们分别建立 6 个模型：成比例优势累计 logit 模型、成比例相邻分类 logit 模型、成比例连续比 logit 模型和不成比例优势累计 logit 模型、不成比例相邻分类 logit 模型、不成比例连续比 logit 模型，并利用参数显著性检验、偏差 D 指标和 AIC 指标比较 6 个模型的优劣（其中 D 和 AIC 都是越小越好）。

表 8.2: 成比例模型摘要 (括号内含两个数字的，分别是标准误、p 值；包含一个数字的是 p 值)

	成比例优势累计	成比例相邻分类	成比例连续比
截距项 1	-1.29(0.14,0.000)	-0.16(0.11,0.145)	-1.23(0.13,0.000)
截距项 2	-0.17(0.14,0.213)	-0.77(0.10,0.000)	-0.93(0.13,0.000)
x1	-0.24(0.09,0.007)	-0.16(0.06,0.007)	-0.19(0.08,0.021)
x2	0.36(0.06,0.000)	0.23(0.04,0.000)	0.32(0.06,0.000)
D	13.80(0.087)	13.72(0.089)	14.20(0.077)
AIC	90.42	90.34	90.82

在成比例的模型中，三个模型的偏差 D 和 AIC 差别不大，在 0.05 显著水平下，协变量系数都显著，并且模型拟合都比较好。相对而言成比例相邻分类 logit 模型比较好（表8.2）。在不成比例模型中，三个模型的偏差 D（在 0.05 显著水平下，模型拟合都比较好）和 AIC 差别不大，并且都存在协变量系数不显著。相对而言的不成比例相邻分类 logit 模型好些（表8.3）。

表 8.3: 成比例模型摘要 (括号内含两个数字的，分别是标准误、p 值；包含一个数字的是 p 值)

	不成比例优势累计	不成比例相邻分类不成比例连续比	
截距项 1	-1.19(0.16,0.000)	0.11(0.20,0.582)	-1.18(0.16,0.000)
截距项 2	-0.26(0.15,0.089)	-1.03(0.19,0.000)	-1.01(0.19,0.000)
x1:1	-0.31(0.11,0.003)	-0.29(0.13,0.027)	-0.31(0.11,0.003)
x1:2	-0.19(0.10,0.067)	-0.03(0.13,0.785)	-0.02(0.13,0.888)
x2:1	0.33(0.07,0.000)	0.14(0.09,0.122)	0.32(0.07,0.000)
x2:2	0.38(0.07,0.000)	0.32(0.08,0.000)	0.31(0.09,0.000)
D	10.97(0.089)	10.64(0.100)	11.10(0.085)
AIC	91.59	91.25	91.72

成比例和不成比例的相邻分类 logit 模型，我们更倾向于前者。因为：(1) 后者存在协变量系数不显著，如 “x1:2” 和 “x2:1”；而前者不存在。(2) 虽然后者的偏差 D 小，但两者的偏差并没有显著差异 (偏差的差为 3.08，p 值为 0.214)，更何况后者多损失了 2 个自由度；(3) 两者的 AIC 差别不大，前者小些。所以，对于表 1，最终建立的模型是成比例相邻分类 logit

模型：

$$\log \frac{\pi_1}{\pi_2} = -0.16 - 0.16x_1 + 0.23x_2$$
$$\log \frac{\pi_2}{\pi_3} = -0.77 - 0.16x_1 + 0.23x_2$$

所以，在给定住房类型情况下，联系其他居民少的住房不满意而不是比较满意、比较满意而不是非常满意的优势是联系其他居民的 0.85（exp(-0.16) 倍。在给定联系其他居民的情况下，住房类型每上升一个类别，所估计的响应变量处在两个相邻类别中较低一个的发生比需要乘以 1.26（exp(0.23)）。总体上来说，联系其他居民多、住房低的住户比联系其他居民少、住房较高的住户更倾向于满意。结论显然与实际相吻合，说明次序响应变量模型是有效。另外从成比例相邻 logit 模型的预测值（表8.4）可以进一步证明模型拟合好。

8.6 参考文献

1. 吴喜之. 复杂数据统计方法. 中国人民大学出版社，2019.
2. Bilder, Loughin. Analysis of Categorical Data with R. CRC Press, 2015.

表 8.4: 预测值 (括号内是真实值)

低 y1	中 y2	高 y3
28.26%(29.68%)	26.35%(24.65%)	45.39%(45.67%)
36.27%(41%)	26.89%(23.97%)	36.84%(35.03%)
44.80%(37.85%)	26.41%(27.12%)	28.79%(35.03%)
23.25%(18.78%)	25.41%(25.97%)	51.34%(55.25%)
30.62%(31.47%)	26.61%(25.90%)	42.77%(42.63%)
38.83%(38.35%)	26.85%(30.97%)	34.32%(30.78%)

3. Thompson. R (and S-PLUS) Manual to Accompany Agresti ' s Categorical Data Analysis (2002) 2nd edition, 2009.
4. Alan Agrestic. Analysis of Ordinal Categories Data(Second Edition). Wiley& Sons, Inc., New Jersey, 2011
5. Snell.A Scaling Procedure for Ordered Categorical Data.Biometric.1964, 20:292-607
6. Clayton, D.G. Some Odds Ratio Statistic for the Analysis of Order Categorical data. Biometric.1974, 61:525-531
7. Peterson,B., F.E. Harell, Jr. Partial Proportional Odds Models for Ordinal Response Variable. Appl. Statist.1990, 39:205-217

8. Yee, T. W. The VGAM Package for Categorical Data Analysis. Journal of Statistical Software. 2010,32:1–34.
9. Mette Madsen. Statistical Analysis of Multiple Contingency Tables. Two Examples. Scandinavian Journal of Statistics. 1976, 3(3):97-106.

第 8 讲 练习

1. 完成实际数据分析的代码。
2. 完成一份数据分析报告。

第 9 讲 分位数回归

分位数回归 (Quantile Regression) 由 Koenker 和 Bassett(1978)。相比均值回归 (Mean Regression)，它可以全面刻画自变量对条件因变量的分布，对异常值有很强的抵抗性。Yu 等 (2003) 对分位数回归的发展做了综述。R 包 quantreg (Koenker 等, 2015) 可以实现分位数回归。

9.1 总体分位数定义

设 Y 是实值随机变量，$F(y) = P(Y \le y)$ 是其分布函数函数，对任意 $\tau \in (0,1)$，则它的 τ 分位数 (无条件) 定义为

$$Q_\tau(Y) = \text{Arginf}\{y \in \mathbb{R}, F(y) \ge \tau\}.$$

若将分布函数 $F(x)$ 的逆定义为 $F_Y^{-1}(\tau) = \inf\{y \in \mathbb{R}, F(y) \ge \tau\}$，则 $Q_\tau(Y) = F_Y^{-1}(\tau)$。这种定义是的分位数具有唯一性。对于样本实现值可以使用 quantile 函数得到。

分位数模拟的代码

```
> y <- rnorm(1000)
> quantile(y, probs = c(0.1, 0.5,0.9))
       10%        50%        90%
-1.29480525  0.01854632  1.27305081
```

分位数的实现可以转化为最小化问题。为了更好的说明这个问题，我们先讨论大家熟悉的均值 (总体) 问题。对于随机变量 Y。其均值 μ 可以通过最小化 $\mathrm{E}[(Y-\theta)^2]$ 实现。因为

$$\begin{aligned}\mathrm{E}[(Y-\theta)^2] &= E[Y^2] - 2E[Y]\theta + \theta^2 \\ &= (\theta - E[Y])^2 + \{E[Y]^2 - (E[Y])^2\} \\ &= (\theta - E[Y])^2 + \mathrm{Var}(Y)\end{aligned}$$

由于第二项 $\mathrm{Var}(Y)$ 是固定的，所以 $\theta = E(Y) = \mu$。这也是为什么均值回归使用均方误差 (Average Squared Deviation) 的原因。换句话说最小二乘得到结果是均值的原因。对于给定的样本 $y_1, y_2, \ldots, y_n$，它的均值可以通过最小化

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$$

实现。

假设 Y 是连续型随机变量，且分布函数是 F，密度函数 f，则分位数可以通过最小化 $E[\rho_\tau(Y-\theta)]$ 实现。因为

$$E[\rho_\tau(Y-\theta)] = \int_{-\infty}^{\theta}(\tau-1)(Y-\theta)f(y)dy + \int_{\theta}^{+\infty}\tau(Y-\theta)f(y)dy$$

其中 $\rho_\tau(t) = \tau t I(t \geq 0) + (\tau-1)tI(t<0)$ 是检验函数 (Check Function)，$I(\bullet)$ 是示性函数。不包含示性函数的检验函数可以表示为

$$\rho_\tau(t) = \begin{cases} \tau t & t \geq 0 \\ (\tau-1)t & t < 0 \end{cases} \tag{9.1}$$

由于

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\int_{-\infty}^{\theta}(\tau-1)(Y-\theta)f(y)dy \\
&\quad=\int_{-\infty}^{\theta}\frac{\partial}{\partial\theta}(\tau-1)(Y-\theta)f(y)dy+\frac{\partial\theta}{\partial\theta}\times(\tau-1)(Y-\theta)f(y)|_{y=\theta} \\
&\quad=(1-\tau)\int_{-\infty}^{\theta}f(y)dy \\
&\quad=(1-\tau)F(\theta) \\
&\frac{\partial}{\partial\theta}\int_{\theta}^{+\infty}\tau(Y-\theta)f(y)dy \\
&\quad=\int_{\theta}^{+\infty}\frac{\partial}{\partial\theta}\tau(Y-\theta)f(y)dy-\tau(Y-\theta)f(y)|_{y=\theta} \\
&\quad=-\tau\int_{\theta}^{+\infty}f(y)dy \\
&\quad=-\tau(1-F(\theta))
\end{aligned}
$$

所以，

$$
\frac{\partial}{\partial\theta}E[\rho_\tau(Y-\theta)]=(1-\tau)F(\theta)-\tau(1-F(\theta))
$$

进而得到：

$$F(\theta) = \tau$$

对于给定的样本 $y_1, y_2, \ldots, y_n$，它的 τ 分位数可以通过最小化

$$\sum_{i=1}^{n} \rho_\tau (y_i - \theta)$$

得到。

9.2 分位数回归

9.2.1 引言

在讲解分位数回归前，我们先直观比较均值回归和分位数回归。以一元均值回归为例说明，$E(Y|X = x) = \beta x$。给定自变量 x，均值回归得到的是随机变量 $Y|X = x$ 的均值。也就是图9.1中的实直线。类似的，分位数回归 $Q_\tau(Y|X = x) = \beta_\tau x$ 得到的是随机变量 $Y|X = x$ 的 τ 分位数 (图9.1中的虚直线)。

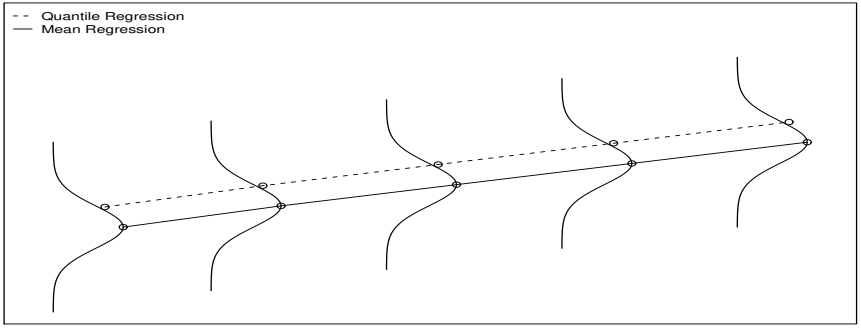


图 9.1: 局部线性回归

R 代码

```
#图1代码
win.graph(width=13,height=6,pointsize=10)
```

```
y <- seq(-4,4,0.1)
x <- dnorm(y)
plot((x+0.1),y,xlim=c(0,4.5), ylim=c(-4,10),xlab="",
    type="l", ylab="", xaxt="n", yaxt="n")
#abline(h=-4)#画竖线
#abline(v=0)#画竖线
points((x+1),(y+1),type="l")
points((x+2),(y+2),type="l")
points((x+3),(y+3),type="l")
points((x+4),(y+4),type="l")
index <- which(x==max(x))
meanx <- x[index] + c(0.1, seq(1:4))
meany <- y[index] + c(0, seq(1:4))
points(meanx,meany, type="o")
points(meanx - dnorm(1.64), meany + 0.95, type="o",lty=2)
legend("topleft",legend=c("Quantile Regression","Mean Regression"),
```

```
lty=c(2,1),cex=0.9,box.lty=0)
```

9.2.2 模型表达

设 $\boldsymbol{Y}=(Y_1,Y_2,\ldots,Y_n)^\top$, $\boldsymbol{X}_i=(X_{i1},X_{i2},\ldots,X_{ip})^\top$ ，线性模型：

$$Y_i=\boldsymbol{X}_i^\top\boldsymbol{\beta}+\varepsilon_i\quad i=1,2,\ldots,n$$

其中 τ 是分位点 (Quantile)，当 $\tau=0.5$ 时，分位数回归是中位数回归。$\boldsymbol{\beta}$ 是 p 维参数向量。ε_i 是随机误差项，且满足 $Q_\tau(\varepsilon_i|\boldsymbol{X}_i)=0$；则分位数回归是

$$Q_\tau(Y_i|\boldsymbol{X}_i=\boldsymbol{x}_i)=\boldsymbol{x}_i^\top\boldsymbol{\beta}_\tau \tag{9.2}$$

其中 $\boldsymbol{x}_i$ 是 $\boldsymbol{X}_i$ 的实现值，$\boldsymbol{\beta}_\tau$ 与 τ 有关。它可以最小化下面式子实现：

$$\sum_i^n\rho_\tau(Y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}_\tau) \tag{9.3}$$

9.2.3 分位数回归求解

我们引入松弛因子 $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)^\top$ 和 $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$，将 (9.3) 式转化为线性规划问题。令

$$Y_i - \boldsymbol{X}_i^\top \beta = u_i - v_i$$

其中 $u_i = \max(0, Y_i - X_i^\top \boldsymbol{\beta})$ 表示正部，$v_i = \max(0, -(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}))$ 表示负部；则

$$\begin{aligned}
\sum_i^n \rho_\tau(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}) &= \sum_{i=1}^n (\tau u_i - (\tau - 1) v_i) \\
&= \tau \boldsymbol{I}_n^\top \boldsymbol{u} + (1 - \tau) \boldsymbol{I}_n^\top \boldsymbol{v} \\
&= (\boldsymbol{0}_{p\times 1}^\top \boldsymbol{\beta} + \tau \boldsymbol{I}_n^\top \boldsymbol{u} + (1 - \tau) \boldsymbol{I}_n^\top \boldsymbol{v} \\
&= (\boldsymbol{0}_{p\times 1}^\top, \tau \boldsymbol{I}_n^\top, (1 - \tau) \boldsymbol{I}_n^\top)(\boldsymbol{\beta}^\top, \boldsymbol{u}^\top, \boldsymbol{v}^\top)^\top \\
&\doteq \mathbf{A}^\top \boldsymbol{\gamma}
\end{aligned}$$

其中 $\boldsymbol{I}_n$ 是 n 维单位向量，$\mathbf{A} = (\boldsymbol{0}_{p\times 1}^\top, \tau \boldsymbol{I}_n^\top, (1 - \tau) \boldsymbol{I}_n^\top)^\top$，$\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{u}^\top, \boldsymbol{v}^\top)^\top$。约束条件是

$$\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{u} - \boldsymbol{v}$$

即

$$\mathbf{X}\boldsymbol{\beta} - \boldsymbol{u} - \boldsymbol{v} = \boldsymbol{Y}$$

从而

$$\begin{bmatrix} \mathbf{X} & \mathbf{E}_n & -\mathbf{E}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \boldsymbol{Y}$$

令 $\mathbf{B} = (\mathbf{X}, \mathbf{E}_n, -\mathbf{E}_n)$，其中 $\mathbf{E}_n$ 是单位阵，则约束条件转化为 $\mathbf{B}\boldsymbol{\gamma} = \boldsymbol{Y}$，从而分位数求解转化为

$$\min \mathbf{A}\boldsymbol{\gamma}$$

$$s.t. \mathbf{B}\boldsymbol{\gamma} = \boldsymbol{Y}$$

R 包 quantreg 中的函数 rq 可以实现。

R 代码

```
rm(list=ls(all=TRUE))#清空所有对象
####产生数据
n <- 100
p <- 2
```

```
tau <- 0.5
x <- rnorm(n)
y <- 1 + 2 * x + rnorm(n)
library(quantreg)
fit.rq <- rq(y~x)
fit.rq$coefficients#rq的解
```

输出结果

```
> fit.rq$coefficients#rq的解
(Intercept)           x
  0.9304538   2.0448851
```

(1) 分位数回归的结果 (参数估计值) 与均值回归的结果没有可比性。因为分位数回归得到的分位数回归线，而均值回归得到均值回归线。一般来说，如果误差项的分布中位数和均值相等 (如标准正态分布) 时，中位数回归与均值回归几乎一样；所以有时文章中会将中位数回归和均值回归相比较。这可以说明数据是否含有异常值。如果两个回归线差距比较大，那么数

据有可能存在异常值。(2) 分位数回归是求解 (9.3) 式。读者注意 $\boldsymbol{\beta}$ 实际与 τ 有关。不同的 τ，得到的不同 $\boldsymbol{\beta}$。高分位点的分位数回归线理论上应该高于低分位点的分位数回归线，但利用 (9.3) 式却不能保证。关于如何解决分位数回归相交的方法，详见 Yu 和 Jones(1998)。

9.2.4 大样本理论

为了简单方便，我们假设 $Y_1, Y_2, \ldots, Y_n$ 是独立的随机变量，并且分布函数分别是 $F_1, F_2, \ldots, F_n$，假设 τ 分位条件函数是

$$Q_\tau(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}_\tau$$

Y_i 的条件分位数函数也可以写成

$$P(Y_i < y|\boldsymbol{X}_i = \boldsymbol{x}_i) = F_{Y_i}(y|\boldsymbol{x}_i) = F_i(y)$$

所以

$$Q_\tau(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = F_{Y_i}^{-1}(\tau|\boldsymbol{x}_i) \equiv \xi_i(\tau).$$

另外，我们记9.3的解为

$$\widehat{\boldsymbol{\beta}}_\tau = \text{argmin}_{\boldsymbol{\beta}_\tau \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau)$$

大样本理论需要下面条件。

C1 $\{F_i\}$ 绝对连续，有连续密度函数 $f_i(\xi)$，它们在点 $\xi_i(\tau)$ 上一致不为 0 和 ∞。

C2 存在正定矩阵 D_0 和 $D_1(\tau)$，使得

(1) $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top = D_0$

(2) $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(\xi_i(\tau)) \boldsymbol{x}_i \boldsymbol{x}_i^\top = D_1(\tau)$

(3) $\max_{i=1,2,\ldots,n} \frac{\|\boldsymbol{x}_i\|}{\sqrt{n}} \to 0$

定理 9.1

在 C1 和 C2 下，

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \sim \mathcal{N}(0, \tau(1-\tau) D_1^{-1} D_0 D_1^{-1})$$

在独立同分布误差模型下，有

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \sim \mathcal{N}(0, \omega^2 D_0^{-1})$$

其中 $\omega^2 = \frac{\tau(1-\tau)}{f_i^2(\xi_i(\tau))}$。证明详见附录。 ♡

sumarry.rq 可以实现对于参数估计值得检验，除了可以实现定理9.1的方法外 (se="nid"、se="iid")，还可以实现秩方法 (Koenker, 1994; se="rank")、核方法 (Powell,1990; se="ker") 和自助法 (Bootstrap, se="boot")。这些方法均在 Koenker(2005) 有详细阐述。其中核方法也是 nid 一种方法。nid 和 ker 区别在于对于 $f_i(\xi_i(\tau))$ 估计方法。秩方法和自助法适合于小样本，尤其是后者。summary.rq 默然秩的方法。需要注意的是秩方法得到秩置信区间，没有参数检验。自助法由于是重复抽样，每一次抽的样本可能不同，所以每一次结果不一样。一般使用 boot 或 nid。

9.2.5 实例分析

例题 **9.1Engel** 数据 Engel 数据包含 235 个观测值，2 个变量：

- income：家庭收入连续变量自变量
- fooexp：食物支出连续变量因变量

该数据主要用来研究家庭支出和家庭收入之间的关系，可以在 quantreg 中 engel 找到。

图9.2中长虚线是均值回归的拟合线，短虚线是中位数回归的拟合线，自下而上的实线分别是分位点为 $\{0.05, 0.1, 0.25, 0.75, 0.9, 0.95\}$ 分位数回归拟合线。随着家庭收入水平的提高，食物支出的增长呈现出扩散的趋势。分位数回归拟合线之间的空间表明食物支出的条件分位数是左偏的 (田茂再,2014)。

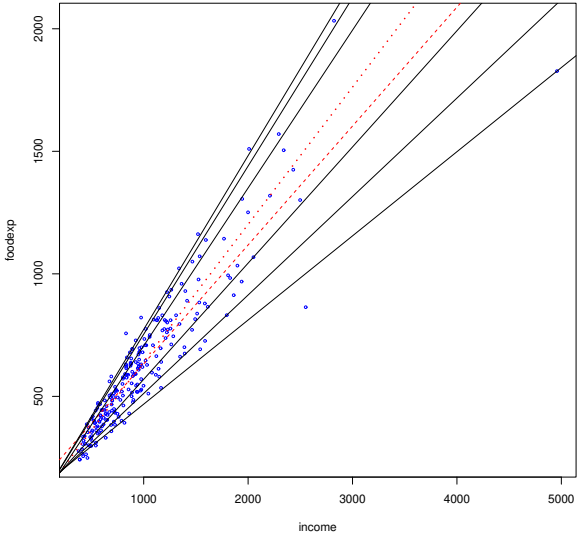


图 9.2: Engle 的拟合曲线

R 代码

```
library(quantreg)
data(engel)
attach(engel)
fit.rq <- rq(foodexp~income, tau=0.5)
summary(fit.rq, se="rank")
summary(fit.rq, se="iid")
summary(fit.rq, se="nid")
summary(fit.rq, se="ker")
summary(fit.rq, se="boot")
#画图
win.graph(width=13, height=13,pointsize=10)
plot(income, foodexp, cex=.25,type="n",xlab="income", ylab="foodexp")
points(income, foodexp, cex=0.5,col="blue")
#中位数回归
abline(rq(foodexp~income, tau=0.5), lwd =2,lty=3, col="red")
```

```
#均值回归
abline(lm(foodexp~income), lty=2, col="red")
taus <- c(0.05, 0.1, 0.25, 0.75, 0.90, 0.95)
for( i in 1:length(taus)) {
  abline(rq(foodexp~income, tau=taus[i])) #分位数回归
}
```

输出结果

```
> summary(fit.rq, se="rank")
Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5
Coefficients:
            coefficients lower bd  upper bd
(Intercept)  81.48225     53.25915 114.01156
income        0.56018      0.48702   0.60199
> summary(fit.rq, se="iid")
```

Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5
Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	81.48225	13.23908	6.15468	0.00000
income	0.56018	0.01192	46.99766	0.00000

> summary(fit.rq, se="nid")
Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5
Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	81.48225	19.25066	4.23270	0.00003
income	0.56018	0.02828	19.81032	0.00000

> summary(fit.rq, se="ker")
Call: rq(formula = foodexp ~ income, tau = 0.5)
tau: [1] 0.5

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	81.48225	30.21532	2.69672	0.00751
income	0.56018	0.03732	15.01139	0.00000

> summary(fit.rq, se="boot")

Call: rq(formula = foodexp ~ income, tau = 0.5)

tau: [1] 0.5

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	81.48225	28.42273	2.86680	0.00453
income	0.56018	0.03771	14.85645	0.00000

9.3 参考文献

1. 田茂再. 复杂数据统计推断理论、方法及应用. 科学出版社, 2014.

2. Hjϕrt N. , Pollard D.(1993). Asymptotics for Minimizers of Convex Processes. Statistical Research Report.
3. Koenker R. Quantile Regression. Cambridge, 2005.
4. Knight K. (1998). Limiting Distributions for L1 Regression Estimators under General Conditions. Annals of Statistics, 26, 755–770.
5. Koenker R., Bassett G. (1978). Regression quantiles. Econometrica, 46, 33–50.
6. Koenker R. 等 (2015). quantreg. R package version 5.19, http://CRAN.R-project.org/package=quantreg.
7. Pollard D. (1991). Asymptotics for Least Absolute Deviation Regression Estimators. Econometric Theory, 7, 186–199.
8. Yu, K., Jones, M. (1998). Local linear quantile regression. Journal of the American statistical Association, 93(441), 228-237.
9. Yu K., Lu Z., Stander J. (2003). Quantile regression: applications and current research areas. Journal of the Royal Statistical Society: Series D (The Statistician), 52, 331–350.
10. Zou H., Yuan M.(2008). Composite quantile regression and the oracle model selection theory. The Annals of Statistics,36(3): 1108–1126.

9.4 附录

10.1的证明。我们考虑下面目标函数

$$Z_n(\boldsymbol{\delta}) = \sum_{i=1}^{n} \left[\rho_\tau(u_i - \boldsymbol{x}_i^\top \frac{\delta}{\sqrt{n}}) - \rho_\tau(u_i) \right]$$

其中 $u_i = Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau$。显然函数 $Z_n(\delta)$ 是凸的 (Convex)，并且在

$$\widehat{\boldsymbol{\delta}}_n = \sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau)$$

取得最小值。根据 Knight(1998) 可以得到 $\widehat{\boldsymbol{\delta}}_n$ 的极限决定于 $Z_n(\boldsymbol{\delta})$ 的极限。根据 Knight 等式，有

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v [I(u \le s) - I(u \le 0)]ds$$

其中 $\psi_\tau(u) = \tau - I(u < 0)$，所以，$Z_n(\boldsymbol{\delta})$ 可以重写为

$$Z_n(\boldsymbol{\delta}) = Z_{1n}(\boldsymbol{\delta}) + Z_{2n}(\boldsymbol{\delta})$$

其中

$$Z_{1n}(\boldsymbol{\delta}) = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{\delta} \psi_\tau(u_i)$$

$$Z_{2n}(\boldsymbol{\delta}) = \sum_{i=1}^{n} \int_0^{v_{ni}} [I(u_i \le s) - I(u_i \le 0)] ds \equiv \sum_{i=1}^{n} Z_{2ni}(\boldsymbol{\delta})$$

其中 $v_{ni} = \frac{\boldsymbol{x}_i^\top \boldsymbol{\delta}}{\sqrt{n}}$。根据 Lindeberg–Feller 中心极限定理和条件 C2，

$$Z_{1n}(\boldsymbol{\delta}) \xrightarrow{\mathrm{d}} -\boldsymbol{\delta}^\top \boldsymbol{W},$$

其中 $\boldsymbol{W} \sim \mathcal{N}(0, \tau(1-\tau)D_0)$。

对于 $Z_{2n}(\boldsymbol{\delta})$，我们有

$$Z_{2n}(\boldsymbol{\delta}) = \sum_{i=1}^{n} \mathrm{E} Z_{2ni}(\boldsymbol{\delta}) + \sum_{i=1}^{n} \left[Z_{2ni}(\boldsymbol{\delta}) - \mathrm{E} Z_{2ni}(\boldsymbol{\delta}) \right]$$

由于

$$
\begin{aligned}
\sum_{i=1}^{n} \mathrm{E} Z_{2ni}(\boldsymbol{\delta}) &= \sum_{i=1}^{n} \int_0^{v_{ni}} \Big[F_i(\xi_i + s) - F_i(\xi_i) \Big] \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_0^{\boldsymbol{x}_i^\top \boldsymbol{\delta}} \Big[F_i(\xi_i + t/sqrtn) - F_i(\xi_i) \Big] dt \\
&= n^{-1} \sum_{i=1}^{n} \int_0^{\boldsymbol{x}_i^\top \boldsymbol{\delta}} \sqrt{n} \Big[F_i(\xi_i + t/sqrtn) - F_i(\xi_i) \Big] dt \\
&= n^{-1} \sum_{i=1}^{n} \int_0^{\boldsymbol{x}_i^\top \boldsymbol{\delta}} f_i(\xi_i) dt + o(1) \\
&= (2n)^{-1} \sum_{i=1}^{n} f_i(\xi_i) \boldsymbol{\delta}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\delta} + o(1) \\
&\to \frac{1}{2} \boldsymbol{\delta}^\top D_1 \boldsymbol{\delta}
\end{aligned}
$$

对于 $Z_{2ni}(\boldsymbol{\delta})$，我们有

$$\begin{aligned}\mathrm{Var}[Z_{2n}(\boldsymbol{\delta})] &= \sum_{i=1}^{n} \mathrm{E}\Big[Z_{2ni}(\boldsymbol{\delta}) - \mathrm{E}[Z_{2ni}(\boldsymbol{\delta})]\Big] \\ &\leq \frac{2}{n}\max_{1\leq i\leq n}|\boldsymbol{x}_i^{\top}\boldsymbol{\delta}| \sum_{i=1}^{n} \mathrm{E}(Z_{2ni})(\boldsymbol{\delta}) \\ &= \frac{2}{n}\max_{1\leq i\leq n}|\boldsymbol{x}_i^{\top}\boldsymbol{\delta}| \sum_{i=1}^{n} \mathrm{E}(Z_{2n})(\boldsymbol{\delta})\end{aligned}$$

所以，如果 $\boldsymbol{\delta}^{\top} D_1 \boldsymbol{\delta} < \infty$，则

$$Z_{2n}(\delta) - \mathrm{E}(Z_{2n}(\delta)) \xrightarrow{\mathrm{p}} 0, n \to \infty$$

从而 $Z_{2n}(\delta) \xrightarrow{\mathrm{p}} \frac{1}{2}\boldsymbol{\delta} D_1 \boldsymbol{\delta}$。如果 $\boldsymbol{\delta} D_1 \boldsymbol{\delta} = \infty$，则

$$
\begin{aligned}
& P\Big(|Z_{2n}(\delta) - \mathrm{E}(Z_{2n}(\delta)) > \varepsilon \mathrm{E}(Z_{2n}(\delta))|\Big) \\
& \leq \frac{\mathrm{Var}(Z_{2n}(\boldsymbol{\delta}))}{\varepsilon^2 \mathrm{E}^2(Z_{2n}(\boldsymbol{\delta}))} \\
& \leq 2\frac{\max_{1\leq i\leq n}|\boldsymbol{x}_i^\top \boldsymbol{\delta}|/\sqrt{n}}{\varepsilon^2 \mathrm{E}^2(Z_{2n}(\boldsymbol{\delta}))} \\
& \to 0
\end{aligned}
$$

$$
Z_{2n}(\delta) - \mathrm{E}(Z_{2n}(\delta)) \xrightarrow{\mathrm{p}} 0, n \to \infty
$$

从而 $Z_{2n}(\delta) \xrightarrow{\mathrm{p}} \infty = \frac{1}{2}\boldsymbol{\delta}^\top D_1 \boldsymbol{\delta}$.

由此，我们可以得到

$$
Z_n(\boldsymbol{\delta}) \xrightarrow{\mathrm{d}} Z_0(\boldsymbol{\delta}) = -\boldsymbol{\delta}^\top \boldsymbol{W} + \frac{1}{2}\boldsymbol{\delta}^\top D_1 \boldsymbol{\delta}
$$

由于 $Z(\delta)$ 有唯一的最小值，所以

$$
\sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) = \widehat{\boldsymbol{\delta}}_n = \mathrm{argmin} Z_n(\delta) \xrightarrow{\mathrm{d}} \widehat{\boldsymbol{\delta}}_0 = \mathrm{argmin} Z_0(\boldsymbol{\delta})
$$

(参见 Pollard(1991)；Hjϕrt 和 Pollard(1993)；Knight(1998))。最后，我们可以看出 $\widehat{\boldsymbol{\delta}}_0 = D^{-1}\boldsymbol{W}$。

由于 $\boldsymbol{W} \sim \mathcal{N}(0, \tau(1-\tau)D_0)$，所以

$$\mathrm{Var}(D_1^{-1}W) = \tau(1-\tau)D_1^{-1}D_0D_1^{-1}$$

从而

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \sim \mathcal{N}(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}).$$

第 10 讲 众数回归

内容提要

- 众数
- 众数回归
- MEM 算法

常见的数字特征主要有均值、分位数和众数[1]。前面我们已经介绍了均值喝分位数。这一讲，我们主要讨论众数。众数是一组数据分布的峰值，对极端值比较稳健；其缺点不唯一[2]。

[1]建议先讲解核密度估计，因为众数回归模型的求解需要使用核密度估计解释

[2]贾俊平, 何晓群, 金勇金. 统计学（第六版）, 中国人民大学,2014.

10.1 模型表达

相比均值回归的均值估计和分位数回归的分位数估计，众数回归是寻找最大可能的估计，即众数的估计。众数回归最早可以追朔到 1989 年 (Lee 1989)。我们考虑如下模型：

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i, i = 1, 2, \ldots, n$$

其中 $\boldsymbol{X}_i \in \mathbb{R}^p$，$\boldsymbol{\beta}_0$ 是属于参数。给定 $\boldsymbol{x}_i$ 的 ε_i 条件密度函数在 $\varepsilon_i = 0$ 有严格全局最大值，也即是 $Mode(Y|X = x) = \boldsymbol{x}^\top \boldsymbol{\beta}_0$。

众数回归是求下面目标函数的最大值。

$$Q_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}}{h_n}\right) \tag{10.1}$$

其中 $K(\bullet)$ 是核函数，通常取高斯核函数（只有选择这个核函数算法才能简化）。h_n 是带宽 (Bandwidth)，与 n 有关。

10.2 MEM 算法

(10.1) 是非线性优化问题的求解，Yao 和 Li(2014) 提出的 MEM (Modal Expectation-maximization) 算法。这种算法主要通过两步实施，类似于 EM(Expectation-maximization) 算法。

E 步 求

$$\pi(i|\boldsymbol{\beta}^k) = \frac{K_{h_n}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}^k)}{\sum_{i=1}^n K_{h_n}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}^k)}$$

M 步

$$\begin{aligned}\boldsymbol{\beta}^{k+1} &= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[\pi(j|\boldsymbol{\beta}^k) \log K_{h_n}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})\right] \\ &= (\boldsymbol{X}^\top \boldsymbol{W}_k \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}_k \boldsymbol{Y}\end{aligned}$$

其中 $W = diag(\pi(1|\boldsymbol{\beta}^k), \pi(2|\boldsymbol{\beta}^k), \ldots, \pi(n|\boldsymbol{\beta}^k))$。

需要注意的是 M 步需要核函数是高斯函数。M 步实际上是加权最小二乘。

10.3 h 的选择

我们现在讨论 h 的选择。我们采用 Yao 等（2012）的方法：

$$F(h) = \frac{1}{n}\sum_{i=1}^{n} K_h''(\hat{\varepsilon}_i)$$

$$G(h) = \frac{1}{n}\sum_{i=1}^{n} \{K_h'(\hat{\varepsilon}_i)\}^2.$$

$\hat{h}_{opt}$ 可以通过最小化

$$\hat{r}(h) = \frac{G(h)F^{-2}(h)}{\hat{\sigma}^2} \tag{10.2}$$

其中 $\hat{\varepsilon}_i = Y_i - \boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}}$，其不需要利用众数回归得到，只要求简单的稳健的方法即可，比如分位数回归。实际中，我们经常采用中位数回归 $h = 0.5\hat{\sigma} \times 1.02^j, j = 0, 1, \ldots, l$。$l = 60$ 或者 100。这里 Here $\hat{\sigma}$ 是 $\hat{\varepsilon}$ 标准差。

笔记 h 的选择与众数回归无关，看似不合理。我们简单讨论一下。

$$\exp(t) \approx 1 + t$$

给定 h_n

$$
\begin{aligned}
Q_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}K\Big(\frac{Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}}{h_n}\Big) \\
&\propto \sum_{i=1}^{n} K\Big(\frac{Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}}{h_n}\Big) \\
&= \sum_{i=1}^{n} \exp\Big(-\frac{(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta})^2}{2h_n^2}\Big) \\
&\approx \sum_{i=1}^{n}\Big(1 - \frac{(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta})^2}{2h_n^2}\Big) \\
&= n - \frac{1}{2h_n^2}(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta})^2 \\
&\propto -\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta})^2
\end{aligned}
$$

我们可以看出近似目标函数 $Q_n(\boldsymbol{\beta}$ 与 h_n 无关。近似说明众数回归与均值的关系。

10.4 模拟实验

我们使用的模型是 Yao 和 Li(2014) 文章中模拟的例子：

$$Y = 1 + 3X + \sigma(X)\varepsilon$$

其中 X 来自 $[0,1]$ 的均匀分布，$\varepsilon \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$，$\sigma = 1 + 2X$。我们可以得到 $E(\varepsilon) = 0$，$Mode(\varepsilon) = 1$，所以 $Model(Y|X) = 2 + 5X$。

模拟的代码

```
rm(list=ls())
library(quantreg)
library(Rfast)
data1 <- function(n){
  x <- cbind(1,runif(n))
  ##genered error begin
  comp <- sample(c(0, 1), size = n, prob = c(0.5, 0.5),
                 replace = T)
  e <- rnorm(n, mean = ifelse(comp == 0, -1, 1),
```

```
10              sd = ifelse(comp == 0, 2.5, 0.5))
11    #plot(density(e))# 看 e 的核密度估计
12
13    sigmax <- 1 + 2 * x[, 2]
14    beta <- c(1, 3)
15    y <- x %*% beta + sigmax * e
16    dat = list(y=y, x=x)
17    return(dat)
18  }
19
20  modreg.MEM <- function(theta, y, x, h, tolset=10^(-6),Intercept=FALSE){
21    n <- dim(x)[1]
22    p <- dim(x)[2]
23    if(Intercept==TRUE) x <- cbind(1, x)
24    tol <- 1
25    iter <- 0
```

```
  #loop begin
  while(tol >tolset){ #if tol <=0.001 stop
    #E-step
    w <- dnorm((y- x %*% theta) / h)
    ww <- c(w/sum(w))
    #M-step
    #W <- matrix(0, n, n)
    #diag(W) <- ww
    thetanew <- Rfast::spdinv(t(x) %*% (ww * x)) %*% t(x) %*% (ww * y)
    cha <- abs(thetanew - theta)
    tol <- sum(cha ^ 2)
    theta <- thetanew
    iter<- iter + 1
    if(iter >=200) break
  }
  res <- y- x %*% theta
```

```
  #loop end
  return(list(iter=iter, theta= theta, tol=tol, res=res))
}


select_h <- function(x, y, Intercept=FALSE){
  if(Intercept==TRUE) x <- cbind(1, x)
  n <- length(y)
  resid0 <- quantreg::rq(y~x-1,tau=0.5)$residuals
  ##################selection h   begin duyue
  rh <- rep(0, 60)
  for (j in 1:60) {
    bdh <- 0.5 * sd(resid0) * 1.02 ^ (j-1)
    gh <- mean( (-dnorm(resid0 / bdh) * resid0 / bdh^3)^2   )
    fh <- mean( dnorm(resid0 / bdh) * resid0^2 / bdh^5 - dnorm(resid0 / bdh) /bdh^3 )
    rh[j] <- gh / fh^2 /(sd(resid0)) ^ 2
```

```
  }
  #plot(rh)
  h <-  0.5 * resid0[(order(rh)[1])] * 1.02 ^ (order(rh)[1])
  return(h)
}

n <- 100
dat <- data1(n)
x <- dat$x
y <- dat$y
h.opt <- select_h(x, y)
theta0 <- coef(quantreg::rq(y~x-1,tau=0.5))
fit <- modreg.MEM(theta=theta0, y=y, x=x, h=h.opt)
fit$theta
```

10.5 参考文献

1. Chen Y., Genovese C., Tibshirani R., Wasserman L. (2016). Nonparametric modal regression. The Annals of Statistics, 在线.
2. Kemp G., Silva, J. (2012). Regression towards the mode. Journal of Econometrics, 170(1), 92-101.
3. Lee M. (1989). Mode regression. Journal of Econometrics, 42, 337–349.
4. Yao, W., Lindsay, B. & Li, R. (2012). Local modal regression. Journal of Nonparametric Statistics, 24(3), 647-663.
5. Yao W., Li, L. (2014). A new regression model: modal linear regression. Scandinavian Journal of Statistics, 41(3), 656-671.

第 11 讲 固定效应和随机效应模型

面板数据分析一直以来是计量经济学中的重要组成部分。

11.1 引言

EmplUK 描述英国 (United Kingdom) 的工人和工资 (Employment, Wages) 从 1976 年到 1984 年 140 个体的情况，共计有 1031 个观测值。需要注意的时，这组数据是不平衡的，因为不同组的观测值不同，也就是说不通的个体的观察值不同。从下面的输出结果可以看出，第 1–103 个体有 7 个观测值，第 104–127 个体有 8 个观测值，其它个体有 9 个观测值。从年份的观察频数也可以看出数据是不平衡的。

R 代码

```
library(plm)
?EmplUK
data("EmplUK",package="plm")
```

```
head(EmplUK)#查看数据前6行
attach(EmplUK)
table(year)
table(firm)
```

输出结果

```
> head(EmplUK)#查看数据前6行
  firm year sector    emp     wage capital    output
1    1 1977      7 5.041 13.1516  0.5894  95.7072
2    1 1978      7 5.600 12.3018  0.6318  97.3569
3    1 1979      7 5.015 12.8395  0.6771  99.6083
4    1 1980      7 4.715 13.8039  0.6171 100.5501
5    1 1981      7 4.093 14.2897  0.5076  99.5581
6    1 1982      7 3.166 14.8681  0.4229  98.6151
> table(year)
1976 1977 1978 1979 1980 1981 1982 1983 1984
```

```
 80  138  140  140  140  140  140   78   35
> table(firm)
 1   2   3   4   5   6   7   8   9  10  11  12  13  14
 7   7   7   7   7   7   7   7   7   7   7   7   7   7
15  16  17  18  19  20  21  22  23  24  25  26  27  28
 7   7   7   7   7   7   7   7   7   7   7   7   7   7
29  30  31  32  33  34  35  36  37  38  39  40  41  42
 7   7   7   7   7   7   7   7   7   7   7   7   7   7
43  44  45  46  47  48  49  50  51  52  53  54  55  56
 7   7   7   7   7   7   7   7   7   7   7   7   7   7
57  58  59  60  61  62  63  64  65  66  67  68  69  70
 7   7   7   7   7   7   7   7   7   7   7   7   7   7
71  72  73  74  75  76  77  78  79  80  81  82  83  84
 7   7   7   7   7   7   7   7   7   7   7   7   7   7
85  86  87  88  89  90  91  92  93  94  95  96  97  98
 7   7   7   7   7   7   7   7   7   7   7   7   7   7
```

```
 99 100 101 102 103 104 105 106 107 108 109 110 111 112
  7   7   7   7   7   8   8   8   8   8   8   8   8   8
113 114 115 116 117 118 119 120 121 122 123 124 125 126
  8   8   8   8   8   8   8   8   8   8   8   8   8   8
127 128 129 130 131 132 133 134 135 136 137 138 139 140
  9   9   9   9   9   9   9   9   9   9   9   9   9   9
```

表11.1摘要 EmplUK 数据的变量的名称、类型等。我们的目的是分析 wage、capital 对 emp 的影响。

表 11.1: EmplUK 数据摘要

变量名	描述	类型
firm	工厂	分类
year	年	分类
sector	the sector of activity	分类
emp	employment	连续
wage	工资	连续
capital	资本	连续
output	输出	连续

11.2 模型介绍和 R 实现

EmplUK 数据是面板数据，它包含个时间和个体两个维度。这一章，我们主要介绍线性面板数据模型：

$$y_{it} = \alpha_{it} + \boldsymbol{\beta}_{it}^{\top}\boldsymbol{x}_{it} + u_{it} \tag{11.1}$$

其中 $\boldsymbol{x}_{it} = (x_{1,it}, \ldots, x_{p,it})^\top$，$\boldsymbol{\beta}_{it} = \beta_{1,it}, \ldots, \beta_{p,it}$；$i = 1, \ldots, n$ 表示个体 (Subject)，如个人、城市、国家等；$t = 1, \ldots, T$ 是时间 (Time)。$\mathrm{E}(u_{it}|\boldsymbol{x}_{it}) = 0$。记 $N = n \times T$。模型11.1是一个非常一般的模型，根据需要，它的变化种类很多。通常存在如下变形：

- 池子模模型 (Pooling Model)，又称无效应模型 (None Effect Model)：

$$y_{it} = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_{it} + u_{it} \tag{11.2}$$

  将所有数据放到一个池子里面“同等”对待，所以称为池子模型；不存在个体和时间效应，所以称为无效应模型，即模型11.1：α_{it} 和 $\boldsymbol{\beta}_{it}$ 不随个体 i 和时间 t 变化

- 固定效应 (Fixed Model)：
  - 个体效应模型 (Individual model)：模型11.1中 $\boldsymbol{\beta}_{it} = \beta_i$ 不随时间 t 变化。

$$y_{it} = \alpha_i + \boldsymbol{\beta}^\top \boldsymbol{x}_{it} + u_{it} \tag{11.3}$$

    这一章的固定效应，我们默认为个体性效应模型.
  - 时间效应模型 (Individual model)：模型11.1中 $\boldsymbol{\beta}_{it} = \beta_t$ 不随个体 i 变化。

$$y_{it} = \alpha_t + \boldsymbol{\beta}^\top \boldsymbol{x}_{it} + u_{it} \tag{11.4}$$

- 随机效应 (Fandom Model)：模型11.1中 $\boldsymbol{\beta}_{it} = \beta + \eta_i$。它的效应是随机的，因为 η_i 不可观测。

我们以 EmplUK 数据为例讲解固定效应和随机效应模型在 R 的实现。

R 代码

```
library(plm)
data("EmplUK",package="plm")
EmplUK <- plm.data(EmplUK, index = c("firm", "year"))
fit.fe <- plm(emp ~ wage + capital, data = EmplUK, model = "within")#固定效应
summary(fit.fe)#
summary(fixef(fit.fe))#查看个体的效应
fit.re <- plm(emp ~ wage + capital, data = EmplUK, model = "random")#随机效应
summary(fit.re)
```

输出结果

```
> summary(fit.fe)#
Oneway (individual) effect Within Model
```

```
Call:
plm(formula = emp ~ wage + capital, data = EmplUK, model = "within")

Unbalanced Panel: n=140, T=7-9, N=1031

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-17.1000  -0.3060   0.0137   0.3070  27.3000

Coefficients :#没有截距项
         Estimate Std. Error t-value  Pr(>|t|)
wage    -0.143626   0.032790 -4.3802 1.327e-05 ***
capital  0.801495   0.064088 12.5062 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Total Sum of Squares:    5030.6
Residual Sum of Squares: 4207.8
R-Squared:      0.16356
Adj. R-Squared: 0.14103
F-statistic: 86.9179 on 2 and 889 DF, p-value: < 2.22e-16
> summary(fixef(fit.fe))#查看个体的效应
   Estimate Std. Error t-value  Pr(>|t|)
1   5.87703    0.93545  6.2826 3.330e-10 ***
2  59.87144    1.45426 41.1697 < 2.2e-16 ***
3  17.07241    1.20391 14.1808 < 2.2e-16 ***
4  21.49019    1.12716 19.0658 < 2.2e-16 ***
5  68.58195    1.85658 36.9400 < 2.2e-16 ***
6   3.90465    1.11012  3.5173 0.0004359 ***
> summary(fit.re)
Oneway (individual) effect Random Effect Model
```

(Swamy-Arora's transformation)

Call:
plm(formula = emp ~ wage + capital, data = EmplUK, model = "random")

Unbalanced Panel: n=140, T=7-9, N=1031

Effects:

	var	std.dev	share
idiosyncratic	4.733	2.176	0.061
individual	72.655	8.524	0.939

theta :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.9040	0.9040	0.9040	0.9064	0.9101	0.9152

Residuals :

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-14.0000  -0.6260  -0.3390  -0.0113   0.0856  29.9000

Coefficients :
             Estimate Std. Error t-value  Pr(>|t|)
(Intercept)  9.072269   1.116282  8.1272 1.253e-15 ***
wage        -0.157677   0.033526 -4.7031 2.912e-06 ***
capital      1.132022   0.059248 19.1066 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    7396.1
Residual Sum of Squares: 5366.2
R-Squared:      0.27448
Adj. R-Squared: 0.27369
F-statistic: 194.44 on 2 and 1028 DF, p-value: < 2.22e-16
```

用 fixef 函数调出固定效应中的个体效应。需要注意的是固定效应没有截距项，因为每一个个体都需要估计，截距项包含着个体效应中提取不出来。随机效应的估计方法 plm 函数可以实现四种，详见 plm 帮助文档。

Hausman 检验主要对比固定效应模型和随机效应模型。

R 代码

```
summary(fit.re)
phtest(fit.fe,fit.re)
```

输出结果

```
Hausman Test
data:  emp ~ wage + capital
chisq = 181.68, df = 2, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

说明拒绝原假设，说明两个模型存在显著差异。

第 12 讲 纵向数据回归

内容提要

- 纵向数据
- 回归

第 13 讲 分层数据回归

内容提要

- 分层数据
- 回归

第 14 讲 超高维变量筛选

第 15 讲 海量数据分析

15.1 线性回归

假设 n 个数据被分成 K 块，记玫