数学建模算法与应用

第11章　偏最小二乘回归分析

在实际问题中，经常遇到需要研究两组多重相关变量间的相互依赖关系，并研究用一组变量（常称为自变量或预测变量）去预测另一组变量（常称为因变量或响应变量），除了最小二乘准则下的经典多元线性回归分析（MLR），提取自变量组主成分的主成分回归分析（PCR）等方法外，还有近年发展起来的偏最小二乘（PLS）回归方法。

偏最小二乘回归提供一种多对多线性回归建模的方法，特别当两组变量的个数很多，且都存在多重相关性，而观测数据的数量（样本量）又较少时，用偏最小二乘回归建立的模型具有传统的经典回归分析等方法所没有的优点。

偏最小二乘回归分析在建模过程中集中了主成分分析，典型相关分析和线性回归分析方法的特点，因此在分析结果中，除了可以提供一个更为合理的回归模型外，还可以同时完成一些类似于主成分分析和典型相关分析的研究内容，提供一些更丰富、深入的信息。

本章介绍偏最小二乘回归分析的建模方法；通过例子从预测角度对所建立的回归模型进行比较。

11.1 偏最小二乘回归分析

考虑 p 个因变量 $y_1, y_2, \cdots, y_p$ 与 m 个自变量 $x_1, x_2, \cdots, x_m$ 的建模问题。偏最小二乘回归的基本作法是首先在自变量集中提出第一成分 u_1（u_1 是 $x_1, \cdots, x_m$ 的线性组合，且尽可能多地提取原自变量集中的变异信息）；同时在因变量集中也提取第一成分 v_1，并要求 u_1 与 v_1 相关程度达到最大。然后建立因变量 $y_1, \cdots, y_p$ 与 u_1 的回归，如果回归方程已达到满意的精度，则算法中止。

否则继续第二对成分的提取，直到能达到满意的精度为止。若最终对自变量集提取r个成分$u_1,u_2,\cdots,u_r$，偏最小二乘回归将通过建立$y_1,\cdots,y_p$与$u_1,u_2,\cdots,u_r$的回归式，然后再表示为$y_1,\cdots,y_p$与原自变量的回归方程式，即偏最小二乘回归方程式。

为了方便起见，不妨假定p个因变量$y_1,\cdots,y_p$与m个自变量$x_1,\cdots,x_m$均为标准化变量。自变量组和因变量组的n次标准化观测数据矩阵分别记为

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix},\quad B=\begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{np} \end{bmatrix}.$$

偏最小二乘回归分析建模的具体步骤如下

（1）分别提取两变量组的第一对成分，并使之相关性达最大。

假设从两组变量分别提出第一对成分为u_1和v_1，u_1是自变量集$X=[x_1,\cdots,x_m]^T$的线性组合

$$u_1=\alpha_{11}x_1+\cdots+\alpha_{1m}x_m=\rho^{(1)T}X,$$

v_1是因变量集$Y=[y_1,\cdots,y_p]^T$的线性组合

$$v_1=\beta_{11}y_1+\cdots+\beta_{1p}y_p=\gamma^{(1)T}Y。$$

为了回归分析的需要，要求

ⅰ）u_1和v_1各自尽可能多地提取所在变量组的变异信息；

ⅱ）u_1和v_1的相关程度达到最大。

由两组变量集的标准化观测数据矩阵A和B，可以计算第一对成分的得分向量，记为$\hat{u}_1$和$\hat{v}_1$

$$\hat{u}_1 = A\rho^{(1)} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} \alpha_{11} \\ \vdots \\ \alpha_{1m} \end{bmatrix}, \tag{11.1}$$

$$\hat{v}_1 = B\gamma^{(1)} = \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{np} \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \vdots \\ \beta_{1p} \end{bmatrix}. \tag{11.2}$$

第一对成分u_1和v_1的协方差$\mathrm{Cov}(u_1, v_1)$可用第一对成分的得分向量$\hat{u}_1$和$\hat{v}_1$的内积来计算。故而以上两个要求可化为数学上的条件极值问题

$$\max \quad (\hat{u}_1 \cdot \hat{v}_1) = (A\rho^{(1)} \cdot B\gamma^{(1)}) = \rho^{(1)T} A^T B\gamma^{(1)}$$

$$\text{s.t.} \begin{cases} \rho^{(1)T}\rho^{(1)} = \left\|\rho^{(1)}\right\|^2 = 1, \\ \gamma^{(1)T}\gamma^{(1)} = \left\|\gamma^{(1)}\right\|^2 = 1. \end{cases} \tag{11.3}$$

利用Lagrange乘数法，问题化为求单位向量$\rho^{(1)}$和$\gamma^{(1)}$，使$\theta_1 = \rho^{(1)T} A^T B \gamma^{(1)}$达到最大。问题的求解只须通过计算$m \times m$矩阵$M = A^T B B^T A$的特征值和特征向量，且$M$的最大特征值为$\theta_1^2$，相应的单位特征向量就是所求的解$\rho^{(1)}$，而$\gamma^{(1)}$可由$\rho^{(1)}$计算得到

$$\gamma^{(1)} = \frac{1}{\theta_1} B^T A \rho^{(1)} \qquad (11.4)$$

(2)建立$y_1,\cdots,y_p$对u_1的回归及$x_1,\cdots,x_m$对u_1的回归。

假定回归模型为

$$\begin{cases} A=\hat{u}_1\sigma^{(1)T}+A_1, \\ B=\hat{u}_1\tau^{(1)T}+B_1, \end{cases} \tag{11.5}$$

其中$\sigma^{(1)}=[\sigma_{11},\cdots,\sigma_{1m}]^T$，$\tau^{(1)}=[\tau_{11},\cdots,\tau_{1p}]^T$分别是多对一的回归模型中的参数向量，$A_1$和$B_1$是残差阵。

回归系数向量$\boldsymbol{\sigma}^{(1)},\boldsymbol{\tau}^{(1)}$的最小二乘估计为

$$\begin{cases}\boldsymbol{\sigma}^{(1)}=\boldsymbol{A}^{T}\hat{\boldsymbol{u}}_1\Big/\left\|\hat{\boldsymbol{u}}_1\right\|^2,\\ \boldsymbol{\tau}^{(1)}=\boldsymbol{B}^{T}\hat{\boldsymbol{u}}_1\Big/\left\|\hat{\boldsymbol{u}}_1\right\|^2,\end{cases} \tag{11.6}$$

称$\boldsymbol{\sigma}^{(1)},\boldsymbol{\tau}^{(1)}$为模型效应负荷量。

（3）用残差阵A_1和B_1代替A和B重复以上步骤。

记$\hat{A}=\hat{u}_1\sigma^{(1)T}$，$\hat{B}=\hat{u}_1\tau^{(1)T}$，则残差阵$E_1=A-\hat{A}$，$B_1=B-\hat{B}$。如果残差阵$B_1$中元素的绝对值近似为0，则认为用第一个成分建立的回归式精度已满足需要了，可以停止抽取成分。否则用残差阵A_1和B_1代替A和B重复以上步骤即得

$$\rho^{(2)}=[\alpha_{21},\cdots,\alpha_{2m}]^T,\quad \gamma^{(2)}=[\beta_{21},\cdots,\beta_{2p}]^T,$$

而$\hat{u}_2 = A_1\rho^{(2)}$，$\hat{v}_2 = B_1\gamma^{(2)}$为第二对成分的得分向量，

$$\sigma^{(2)} = A_1^T\hat{u}_2 \Big/ \left\|\hat{u}_2\right\|^2, \quad \tau^{(2)} = B_1^T\hat{u}_2 \Big/ \left\|\hat{u}_2\right\|^2$$

分别为X,Y的第二对成分的负荷量。这时有

$$\begin{cases} A = \hat{u}_1\sigma^{(1)T} + \hat{u}_2\sigma^{(2)T} + A_2, \\ B = \hat{u}_1\tau^{(1)T} + \hat{u}_2\tau^{(2)T} + B_2. \end{cases}$$

（4）设$n\times m$数据阵A的秩为$r\le \min(n-1,m)$，则存在r个成分$u_1,u_2,\cdots,u_r$，使得

$$\begin{cases} A=\hat{u}_1\sigma^{(1)T}+\cdots+\hat{u}_r\sigma^{(r)T}+A_r, \\ B=\hat{u}_1\tau^{(1)T}+\cdots+\hat{u}_r\tau^{(r)T}+B_r. \end{cases} \tag{11.7}$$

把 $u_k=\alpha_{k1}x_1+\cdots+\alpha_{km}x_m$ （$k=1,2,\cdots,r$），代入 $Y=u_1\tau^{(1)}+\cdots+u_r\tau^{(r)}$，即得$p$个因变量的偏最小二乘回归方程式

$$y_j=c_{j1}x_1+\cdots+c_{jm}x_m,\quad j=1,2,\cdots,p. \tag{11.8}$$

（5）交叉有效性检验。

一般情况下，偏最小二乘法并不需要选用存在的r个成分$u_1,u_2,\cdots,u_r$来建立回归式，而像主成分分析一样，只选用前l个成分（$l \le r$），即可得到预测能力较好的回归模型。对于建模所需提取的成分个数l，可以通过交叉有效性检验来确定。

每次舍去第i个观测数据（$i=1,2,\cdots,n$），对余下的$n-1$个观测数据用偏最小二乘回归方法建模，并考虑抽取h（$h\leq r$）个成分后拟合的回归式，然后把舍去的自变量组第i个观测数据代入所拟合的回归方程式，得到$y_j(j=1,2,\cdots,p)$在第i个观测点上的预测值$\hat{b}_{(i)j}(h)$。

对$i=1,\ 2,\cdots,n$重复以上的验证，即得抽取h个成分时第j个因变量$y_j(j=1,2,\cdots,p)$的预测误差平方和为

$$\mathrm{PRESS}_j(h)=\sum_{i=1}^{n}(b_{ij}-\hat{b}_{(i)j}(h))^2,\quad j=1,2,\cdots,p,$$

$Y=[y_1,\cdots,y_p]^T$的预测误差平方和为

$$\mathrm{PRESS}(h)=\sum_{i=1}^{p}\mathrm{PRESS}_j(h).$$

另外，再采用所有的样本点，拟合含h个成分的回归方程。这时，记第i个样本点的预测值为$\hat{b}_{ij}(h)$，则可以定义y_j的误差平方和为

$$\mathrm{SS}_j(h)=\sum_{i=1}^{n}(b_{ij}-\hat{b}_{ij}(h))^2,$$

定义 的误差平方和为

$$\mathrm{SS}(h)=\sum_{j=1}^{p}\mathrm{SS}_j(h).$$

当PRESS(h)达到最小值时，对应的h即为所求的成分个数l。通常，总有PRESS(h)大于SS(h)，而SS(h)则小于SS($h-1$)。因此，在提取成分时，总希望比值PRESS(h)/SS($h-1$)越小越好；一般可设定限制值为0.05，即当

$$\text{PRESS}(h)/\text{SS}(h-1) \le (1-0.05)^2 = 0.95^2$$

时，增加成分u_h有利于模型精度的提高。

或者反过来说，当

$$\mathbf{PRESS}(h)/\mathbf{SS}(h-1) > 0.95^2$$

时，就认为增加新的成分u_h，对减少方程的预测误差无明显的改善作用。

为此，定义交叉有效性为

$$Q_h^2 = 1 - \mathrm{PRESS}(h)/\mathrm{SS}(h-1),$$

这样，在建模的每一步计算结束前，均进行交叉有效性检验，如果在第h步有$Q_h^2 < 1 - 0.95^2 = 0.0975$，则模型达到精度要求，可停止提取成分；若$Q_h^2 \geq 0.0975$，表示第$h$步提取的$u_h$成分的边际贡献显著，应继续第$h+1$步计算。

11.2 Matlab 偏最小二乘回归命令 plsregress

Matlab 工具箱中偏最小二乘回归命令 plsregress 的使用格式为

[XL,YL,XS,YS,BETA,PCTVAR,MSE,stats] = plsregress (X,Y,ncomp)

其中 X 为 n×m 的自变量数据矩阵，每一行对应一个观测，每一列对应一个变量；Y 为 n×p 的因变量数据矩阵，每一行对应一个观测，每一列对应一个变量；ncomp 为成分的个数，ncomp 的默认值为 min(n-1,m)。返回值 XL 为对应于$\hat{\sigma}_i$的 m×ncomp 的负荷量矩阵，它的每一行为对应于式（11.7）的第一式的回归表达式；YL 为对应于$\hat{\tau}_i$的 p×ncomp 矩阵，它的每一行为对应于式（11.7）的第二式的回归表达式；

XS 是对应于$\hat{\boldsymbol{u}}_i$的得分矩阵，Matlab 工具箱中对应于式（11.3）的特征向量$\boldsymbol{\rho}^{(i)}$不是取为单位向量，$\boldsymbol{\rho}^{(i)}$取为使得每个$\hat{\boldsymbol{u}}_i$对应的得分向量是单位向量，且不同的得分向量是正交的；YS 是对应于$\hat{\boldsymbol{v}}_i$的得分矩阵，它的每一列不是单位向量，列与列之间也不正交；BETA 的每一列为对应于式（11.8）的回归表达式；PCTVAR 是一个两行的矩阵，第一行的每个元素对应着自变量提出成分的贡献率，第二行的每个元素对应着因变量提出成分的贡献率；

MSE 是一个两行的矩阵，第一行的第j个元素对应着自变量与它的前$j-1$个提出成分之间回归方程的剩余标准差，第二行的第j元素对应着因变量与它的前$j-1$个提出成分之间回归方程的剩余标准差；stats 返回 4 个值，其中返回值 stats.W 的每一列对应着特征向量$\rho^{(i)}$，这里的特征向量不是单位向量。

11.3　案例分析

例 11.1　本例采用兰纳胡德（Linnerud）给出的关于体能训练的数据进行偏最小二乘回归建模。在这个数据系统中被测的样本点，是某健身俱乐部的 20 位中年男子。被测变量分为两组。第一组是身体特征指标 X，包括体重、腰围、脉搏。第二组变量是训练结果指标Y，包括单杠、弯曲、跳高。原始数据见表 11.1(表略)。

解　x_1, x_2, x_3分别表示自变量指标体重、腰围、脉搏，y_1, y_2, y_3分别表示因变量指标单杠、弯曲、跳高，自变量的观测数据矩阵记为$A = (a_{ij})_{20\times3}$，因变量的观测数据矩阵记为$B = (b_{ij})_{20\times3}$。

（1）数据标准化

将各指标值a_{ij}转换成标准化指标值$\tilde{a}_{ij}$，

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j^{(1)}}{s_j^{(1)}},\quad i=1,2,\cdots,20,\quad j=1,2,3,$$

其中$\mu_j^{(1)}=\frac{1}{20}\sum_{i=1}^{20}a_{ij}$，$s_j^{(1)}=\sqrt{\frac{1}{20-1}\sum_{i=1}^{n}(a_{ij}-\mu_j^{(1)})^2}$，

$(j=1,2,3)$，即$\mu_j^{(1)},s_j^{(1)}$为第j个自变量x_j的样本均值和样本标准差。对应地，称

$$\tilde{x}_j=\frac{x_j-\mu_j^{(1)}}{s_j^{(1)}},\quad j=1,2,3,$$

为标准化指标变量。

类似地，将b_{ij}转换成标准化指标值$\tilde{b}_{ij}$，

$$\tilde{b}_{ij}=\frac{b_{ij}-\mu_j^{(2)}}{s_j^{(2)}},\ i=1,2,\cdots,20,\ j=1,2,3,$$

其中$\mu_j^{(2)}=\frac{1}{20}\sum_{i=1}^{20}b_{ij}$，$s_j^{(2)}=\sqrt{\frac{1}{20-1}\sum_{i=1}^{n}(b_{ij}-\mu_j^{(2)})^2}$，

$(j=1,2,3)$，即$\mu_j^{(2)},s_j^{(2)}$为第j个因变量y_j的样本均值和样本标准差；对应地，称

$$\tilde{y}_j=\frac{y_j-\mu_j^{(2)}}{s_j^{(2)}},\ j=1,2,3,$$

为对应的标准化变量。

（2）求相关系数矩阵

表 11.2 给出了这 6 个变量的简单相关系数矩阵。从相关系数矩阵可以看出，体重与腰围是正相关的；体重、腰围与脉搏负相关；而在单杠、弯曲与跳高之间是正相关的。从两组变量间的关系看，单杠、弯曲和跳高的训练成绩与体重、腰围负相关，与脉搏正相关。

表 11.2　相关系数矩阵

	体重 (x_1)	腰围 (x_2)	脉搏 (x_3)	单杠 (y_1)	弯曲 (y_2)	跳高 (y_3)
体重(x_1)	1	0.8702	-0.3658	-0.3897	-0.4931	-0.2263
腰围(x_2)	0.8702	1	-0.3529	-0.5522	-0.6456	-0.1915
脉搏(x_3)	-0.3658	-0.3529	1	0.1506	0.225	0.0349
单杠(y_1)	-0.3897	-0.5522	0.1506	1	0.6957	0.4958
弯曲(y_2)	-0.4931	-0.6456	0.225	0.6957	1	0.6692
跳高(y_3)	-0.2263	-0.1915	0.0349	0.4958	0.6692	1

（3）分别提出自变量组和因变量组的成分

使用 Matlab 软件，求得的各对成分分别为

$$\begin{cases} u_1 = -0.0951\tilde{x}_1 - 0.1244\tilde{x}_2 + 0.0385\tilde{x}_3, \\ v_1 = 2.1191\tilde{y}_1 + 2.5809\tilde{y}_2 + 0.8869\tilde{y}_3, \end{cases}$$

$$\begin{cases} u_2 = -0.1279\tilde{x}_1 + 0.2429\tilde{x}_2 + 0.2202\tilde{x}_3, \\ v_2 = -0.8054\tilde{y}_1 - 0.1171\tilde{y}_2 - 0.5486\tilde{y}_3, \end{cases}$$

$$\begin{cases} u_3 = -0.4416\tilde{x}_1 + 0.3790\tilde{x}_2 - 0.1055\tilde{x}_3, \\ v_3 = -0.7781\tilde{y}_1 - 0.1987\tilde{y}_2 + 0.0381\tilde{y}_3. \end{cases}$$

前两个成分解释自变量的比率为 92.13%，只要取两对成分即可。

（4）求两个成分对时标准化指标变量与成分变量之间的回归方程

求得自变量组和因变量组与u_1, u_2之间的回归方程分别为

$$\tilde{x}_1 = -4.1306u_1 + 0.0558u_2,$$
$$\tilde{x}_2 = -4.1933u_1 + 1.0239u_2,$$
$$\tilde{x}_3 = 2.2264u_1 + 3.4441u_2,$$
$$\tilde{y}_1 = 2.1191u_1 - 0.9714u_2,$$
$$\tilde{y}_2 = 2.5809u_1 - 0.8398u_2,$$
$$\tilde{y}_3 = 0.8869u_1 - 0.1877u_2.$$

（5）求因变量组与自变量组之间的回归方程

把（3）中成分u_i代入（4）中$\tilde{y}_i$的回归方程，得到标准化指标变量之间的回归方程

$$\tilde{y}_1 = -0.0773\tilde{x}_1 - 0.4995\tilde{x}_2 - 0.1323\tilde{x}_3,$$

$$\tilde{y}_2 = -0.1380\tilde{x}_1 - 0.5250\tilde{x}_2 - 0.0855\tilde{x}_3,$$

$$\tilde{y}_3 = -0.0603\tilde{x}_1 - 0.1559\tilde{x}_2 - 0.0072\tilde{x}_3.$$

将标准化变量$\tilde{y}_j, \tilde{x}_j (j=1,2,3)$分别还原成原始变量$y_j, x_j$，得到回归方程

$$y_1 = 47.0375 - 0.0165x_1 - 0.8246x_2 - 0.0970x_3,$$

$$y_2 = 612.7674 - 0.3497x_1 - 10.2576x_2 - 0.7422x_3,$$

$$y_3 = 183.9130 - 0.1253x_1 - 2.4964x_2 - 0.0510x_3.$$

（6）模型的解释与检验

为了更直观、迅速地观察各个自变量在解释$y_j(j=1,2,3)$时的边际作用，可以绘制回归系数图，见图 11.1。这个图是针对标准化数据的回归方程的。

从回归系数图中可以立刻观察到，腰围变量在解释三个回归方程时起到了极为重要的作用。然而，与单杠及弯曲相比，跳高成绩的回归方程显然不够理想，三个自变量对它的解释能力均很低。

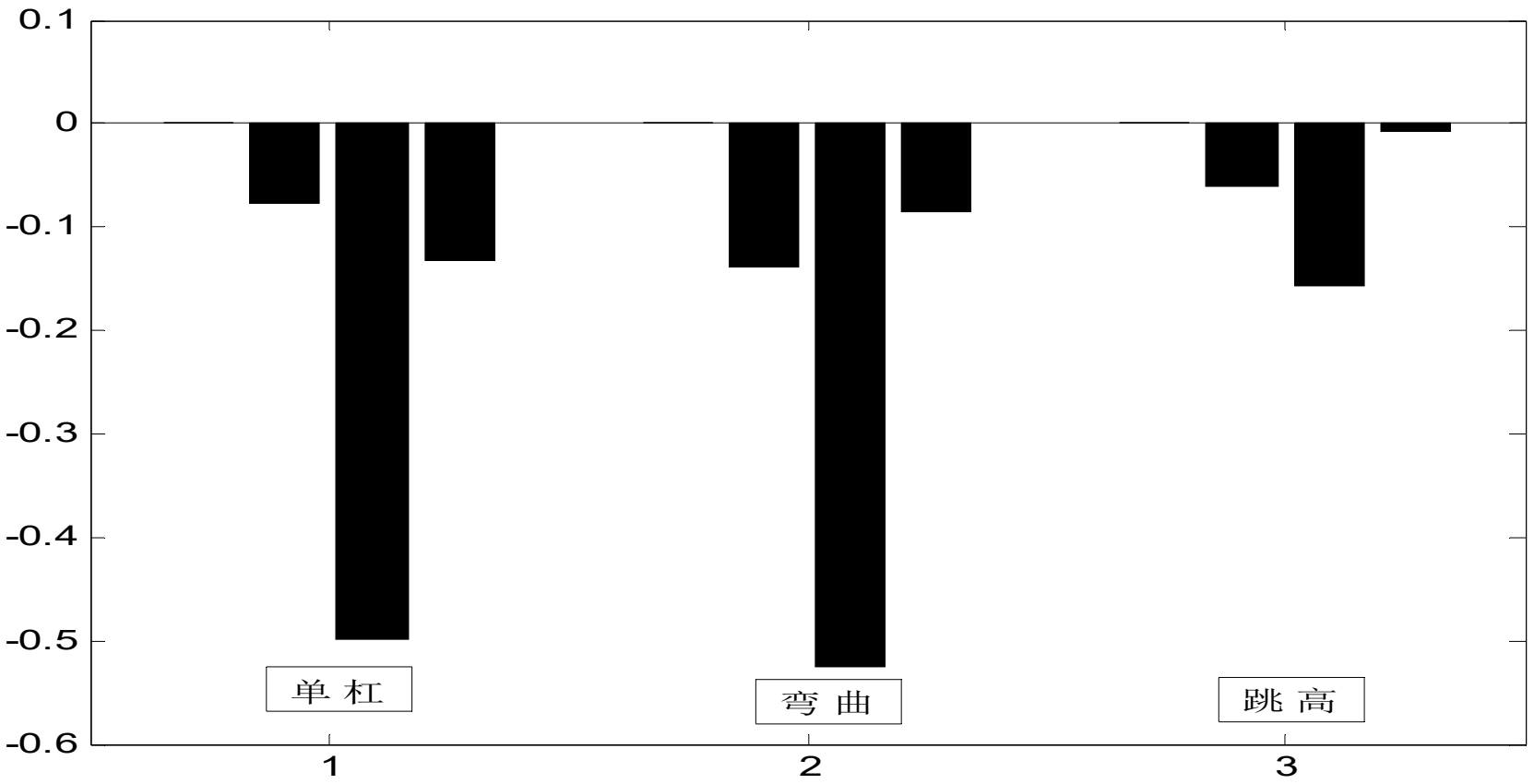

图 11.1　回归系数的直方图

为了考察这三个回归方程的模型精度，我们以$(\hat{y}_{ij}, y_{ij})$为坐标值，对所有的样本点绘制预测图。$\hat{y}_{ij}$是第j个因变量指标在第i个样本点(y_{ij})的预测值。在这个预测图上，如果所有点都能在图的对角线附近均匀分布，则方程的拟合值与原值差异很小，这个方程的拟合效果就是满意的。体能训练的预测图见图 11.2。

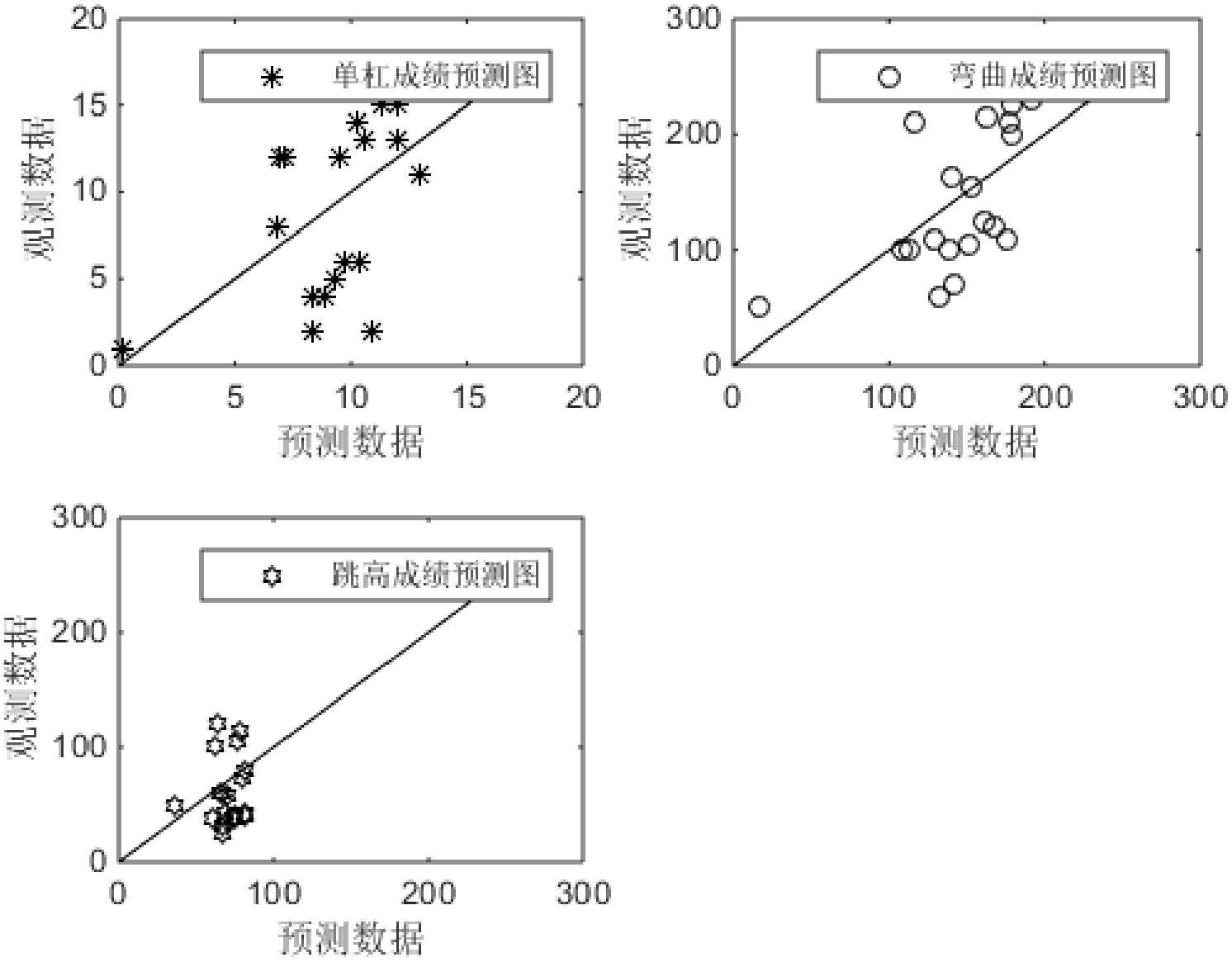

图 11.2　体能训练预测图

例 11.2　交通运输业和旅游业是相关行业，两者之间存在密切的关系。一方面，旅游业是综合产业，它的发展会带动交通运输等产业的发展，交通客运的客源主力正是旅游者。另一方面，交通运输业对旅游业有着重要影响。第一，交通运输是发展旅游业的前提和命脉。交通运输作为旅游业“行、游、住、食、购、娱”六要素中的“行”，是旅游业发展的硬件基础，旅游地只有注重交通运输建设，具备良好的可进入性，旅游人数才会逐年增加，旅游业才能得到发展。

第二，交通运输是旅游业中旅游收入和旅游创汇的重要来源。第三，交通运输业影响旅游者的旅游意愿。交通运输业的发展状况、价格、服务质量、便利程度等都会影响人们的旅游意愿，从而影响旅游业的发展。交通运输的建设布局和运力投入，可以调节旅游业的发展规模。但旅游业与交通运输业存在着相辅相成、相互制约的关系。交通的阻塞问题已经成为旅游业发展的瓶颈。

为研究交通运输业与旅游业之间的关系，我们选择了客运量指标及旅游业相关指标。客运量指标选择了铁路客运量y_1、公路客运量y_2、水运客运量y_3和民航客运量y_4四个指标。为反映旅游业的发展情况我们选择了旅行社数x_1（个）、旅行社从业人员x_2（人）、入境旅游人数x_3（万人次）、国内居民出境人数x_4（万人次）、国内旅游人数x_5（亿人次）、国际旅游外汇收入x_6（亿美元）和国内旅游收入（亿元）等七个指标。指标数据见表 11.3(表略)，来源于《中国统计年鉴》，数据区间为 1996—2006 年。拟运用偏最小二乘法分析这些变量之间的关系。

解　（1）数据标准化

这里数据的量纲和数量级差异很大，首先进行数据标准化。

（2）建立偏最小二乘回归模型

利用 Matlab 软件的计算结果，可以发现，最终只需选取前两对成分，对自变量组的解释比率为 98.55%，对因变量组的解释比率为 72.64%。这说明效果是不错的。

标准化变量的偏最小二乘回归方程为

$$\tilde{y}_1 = 0.0103\tilde{x}_1 - 0.1019\tilde{x}_2 - 0.0034\tilde{x}_3 + 0.2559\tilde{x}_4 + 0.3404\tilde{x}_5 + 0.2785\tilde{x}_6 + 0.1607\tilde{x}_7$$

$$\tilde{y}_2 = -0.2845\tilde{x}_1 - 0.4648\tilde{x}_2 - 0.3146\tilde{x}_3 + 0.1645\tilde{x}_4 + 0.3065\tilde{x}_5 + 0.1885\tilde{x}_6 - 0.0262\tilde{x}_7$$

$$\tilde{y}_3 = -0.6418\tilde{x}_1 - 1.1426\tilde{x}_2 - 0.7213\tilde{x}_3 + 0.5789\tilde{x}_4 + 0.9702\tilde{x}_5 + 0.6517\tilde{x}_6 + 0.0677\tilde{x}_7$$

$$\tilde{y}_4 = 0.0034\tilde{x}_1 - 0.1211\tilde{x}_2 - 0.0121\tilde{x}_3 + 0.2774\tilde{x}_4 + 0.3713\tilde{x}_5 + 0.3022\tilde{x}_6 + 0.1708\tilde{x}_7$$

最终得到的偏最小二乘回归方程为

$$y_1 = 79424.4109 + 0.0218x_1 - 0.0136x_2 - 0.0139x_3 + 2.5$$
$$+ 1363.7331x_5 + 36.6921x_6 + 1.1572x_7$$

$$y_2 = 129900688.8451 - 5414.9287x_1 - 557.4809x_2$$
$$- 11510.3590x_3 + 14707.6028x_4 + 11070274.6617x_5$$
$$+ 223847.8474x_6 - 1700.6600x_7$$

$$y_3 = 21077.0402 - 0.2465x_1 - 0.0277x_2 - 0.5328x_3$$
$$+ 1.0447x_4 + 707.4398x_5 + 15.6215x_6 + 0.0887x_7$$

$$y_4 = -767.8584 + 0.0026x_1 - 0.0058x_2 - 0.0177x_3$$
$$+ 0.9929x_4 + 536.9458x_5 + 14.3677x_6 + 0.4438x_7$$