第5章

回归模型原理与应用

本章学习目标

- 熟悉线性回归模型的形式。
- 了解线性回归方程参数求解的方法。
- 了解非线性回归模型。
- 掌握线性回归方程的选择和预测。
- 掌握线性回归模型的程序实现。

本章主要内容如图 5.1 所示。

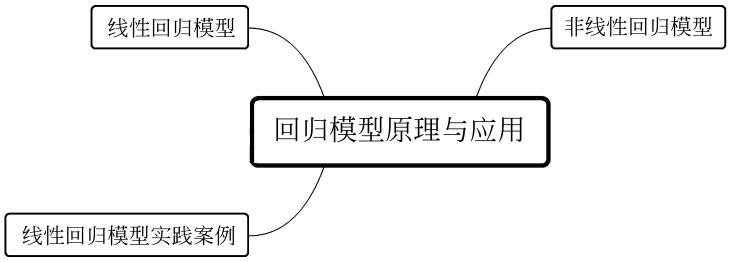


图 5.1　本章主要内容

5.1　线性回归模型

5.1.1　回归分析的含义

“回归”这一术语最早是由英国统计学家高尔顿(Francis Galton)在 19 世纪末期研究人类遗传问题时提出的。他发现,总体上子辈身高和父辈身高是正相关的,但是当父辈身高异常高或异常矮时,子辈高于或矮于父辈的概率很小。他将子辈身高向父辈的平均身高回归的趋势移动称为回归效应。人们更关注给定父辈身高的情形下如何找到子辈平均身高的变化规律。

假设已知一组身高的实测数据,绘制成散点图,如图 5.2 所示。

在图 5.2 中，父亲身高分为 4 组；对应于设定的父亲身高，儿子身高有一个分布范围。

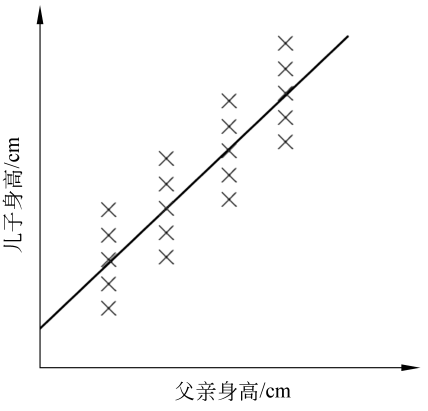


图 5.2　父亲和儿子身高的散点图

如果画一条表示儿子平均身高的线，说明儿子的平均身高是如何随着父亲身高的增加而增加的，这条线就是回归线。从图形上看，回归线是对实测数据的拟合，既可以是直线，也可以是曲线，通过回归线可以预测儿子的身高。从函数形式上看，回归线对应的是函数表达式，这个表达式描述了儿子身高和父亲身高的依赖关系。如果是线性表达式，回归线就是直线；如果是非线性表达式，回归线就是曲线。因此，回归分析的任务就是：使用输入变量 X 建立某个函数关系式，尽可能准确地预测输出变量 Y。按照输入变量的个数，可将回归分析分为一元回归分析和多元回归分析；按照输入变量和输出变量之间的关系类型，可将回归分析分为线性回归分析和非线性回归分析。回归分析的步骤如下：

(1) 建立输出变量与输入变量的回归方程。

(2) 选择最优回归方程。

① 计算回归方程的拟合优度。

② 检验回归方程的整体显著性。

③ 检验各回归系数的显著性。

④ 回归诊断。

(3) 利用最优回归方程进行预测。

5.1.2　线性回归模型的形式

一元线性回归模型采用一元线性回归方程的形式，其公式如下：

$$Y=\theta_0+\theta_1 x_1+\varepsilon$$

其中，Y 是输出变量；θ_0、θ_1 是未知参数；x_1 是输入变量；ε 是误差项，用于反映输入变量和输出变量之间除线性关系之外的随机因素或不可观测的因素，假定 $\varepsilon\sim N(0,\sigma^2)$ 且相互独立。例如，影响儿子身高的因素除了父亲身高之外还有其他随机因素，输入变量只有父亲身高，所以可用该模型表达二者的函数关系。

大多数情况下输出变量会受多个输入变量的影响，例如决定儿子身高的因素不只有父亲身高，假设还有运动时间、饮食习惯等多个因素，即多个输入变量，因此可以建立更一般的包含 k 个输入变量的线性回归方程如下：

$$Y=\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots\theta_k x_k+\varepsilon$$

在实际建模中，输入变量 X 和输出变量 Y 的实测值(即训练数据)是已知的，而参数 $\theta_0,\theta_1,\theta_2,\cdots,\theta_k$ 是未知的，建立线性回归方程的关键是用训练数据求得参数 $\theta_0,\theta_1,\theta_2,\cdots,\theta_k$ 的最优解 $\hat{\theta}_0,\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$，代入上式后略去误差项，得到最终用于预测的线性回归方程：

$$\hat{Y}=\hat{\theta}_0+\hat{\theta}_1x_1+\hat{\theta}_2x_2+\cdots+\hat{\theta}_kx_k$$

5.1.3 线性回归方程参数求解

$\hat{\theta}$随着参数$\theta_0,\theta_1,\theta_2,\cdots,\theta_k$取值的不同，会产生不同的线性回归方程，回归分析的任务就是要从中找到最优方程，即对训练数据拟合获取最优回归线，因此问题就转换成了如何求得参数$\theta_0,\theta_1,\theta_2,\cdots,\theta_k$的最优解$\hat{\theta}_0,\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$，常用的方法有正规方程法和梯度下降法，这两种方法都包含最小二乘法的思想。以一元线性回归为例，最小二乘法的思想是：首先获得n关于输入变量X和输出变量Y的实测值$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，并在坐标轴上绘制$n$对数据点；然后用一条直线拟合这些数据点，最优的拟合直线应该是距离大部分数据点最近的那条直线，即把n对数据点代入最优直线对应的线性回归方程得到的预测值与实测值的差异最小。两者的关系用如下公式表示：

$$Y=\hat{Y}+\hat{\varepsilon}$$

将n对实测值$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$代入上式可得

$$y_i=\hat{y}_i+\hat{\varepsilon}_i,\quad i=1,2,\cdots,n$$

$$\hat{\varepsilon}_i=y_i-\hat{y}_i,\quad i=1,2,\cdots,n$$

其中，y_i称为实测值，$\hat{y}_i$称为预测值；$\hat{\varepsilon}_i$是实测值与其预测值之差，称为残差，也是真实误差ε的估计。实践中一般希望预测值尽可能地靠近实测值，使损失函数尽可能小。

$$J(\hat{\boldsymbol{\theta}})=J(\hat{\theta}_0,\hat{\theta}_1)=\sum\hat{\varepsilon}_i^2=\sum(y_i-\hat{y}_i)^2=\sum(y_i-\hat{\theta}_0-\hat{\theta}_1x_i)^2$$

上式常称为损失函数，其中$\hat{\varepsilon}_i^2$是残差的平方。问题进一步转换成了求解参数向量$\hat{\boldsymbol{\theta}}$，使$J(\hat{\boldsymbol{\theta}})$达到最小。

1. 正规方程法

正规方程法可以一次性求解参数向量$\hat{\boldsymbol{\theta}}$，由微积分知识可知，$J(\hat{\boldsymbol{\theta}})$对$\hat{\theta}_0$和$\hat{\theta}_1$的偏导为0时$J(\hat{\boldsymbol{\theta}})$最小。以一元线性回归方程为例，只需计算$J(\hat{\boldsymbol{\theta}})$对$\hat{\theta}_0$和$\hat{\theta}_1$的偏导，令偏导数为0，即可求得对应的$\hat{\theta}_0$和$\hat{\theta}_1$，计算过程如下：

对$\hat{\theta}_0$、$\hat{\theta}_1$求偏导，可得：

$$\frac{\partial}{\partial\hat{\theta}_0}J(\hat{\boldsymbol{\theta}})=-2\sum(y_i-\hat{\theta}_0-\hat{\theta}_1x_i)$$

$$\frac{\partial}{\partial\hat{\theta}_1}J(\hat{\boldsymbol{\theta}})=-2\sum(y_i-\hat{\theta}_0-\hat{\theta}_1x_i)x_i$$

令$\frac{\partial}{\partial\hat{\theta}_0}J(\hat{\boldsymbol{\theta}})=0$，$\frac{\partial}{\partial\hat{\theta}_1}J(\hat{\boldsymbol{\theta}})=0$，可得：

$$\begin{cases}\sum y_i=n\hat{\theta}_0+\hat{\theta}_1\sum x_i\\ \sum x_iy_i=\hat{\theta}_0\sum x_i+\hat{\theta}_1\sum x_i^2\end{cases}$$

求解联立方程，可得

$$\begin{cases}\hat{\theta}_1=\dfrac{\sum x_i y_i-\dfrac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sum x_i^2-\dfrac{1}{n}\left(\sum x_i\right)^2}=\dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\\ \hat{\theta}_0=\bar{y}-\hat{\theta}_1\bar{x}\end{cases}$$

其中，$\bar{x}=\dfrac{\sum x}{n}$，$\bar{y}=\dfrac{\sum y}{n}$ 分别为 X 和 Y 的样本均值。当给出 θ_0、θ_1 的估计 $\hat{\theta}_0$、$\hat{\theta}_1$ 后，将其代入回归方程，可由 X 求出 Y 的预测值。

在多元线性回归方程中，如果有 k 个输入变量和 n 次独立的样本实测数据 x_{i1}，x_{i2}，…，x_{ik}；y_i），$i=1,2,\cdots,n$，线性回归模型可表示成如下形式：

$$\begin{cases}y_1=\hat{\theta}_0+\hat{\theta}_1x_{11}+\hat{\theta}_2x_{12}+\cdots+\hat{\theta}_kx_{1k}+\hat{\varepsilon}_1\\ y_2=\hat{\theta}_0+\hat{\theta}_1x_{21}+\hat{\theta}_2x_{22}+\cdots+\hat{\theta}_kx_{2k}+\hat{\varepsilon}_2\\ \vdots\\ y_n=\hat{\theta}_0+\hat{\theta}_1x_{n1}+\hat{\theta}_2x_{n2}+\cdots+\hat{\theta}_kx_{nk}+\hat{\varepsilon}_n\end{cases}$$

上式可进一步简写成如下形式：

$$\boldsymbol{Y}=\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}}=\boldsymbol{X}\hat{\boldsymbol{\theta}}+\hat{\boldsymbol{\varepsilon}}$$

其中：

$$\boldsymbol{Y}=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},\quad \hat{\boldsymbol{Y}}_1=\begin{bmatrix}\hat{y}_1\\\hat{y}_2\\\vdots\\\hat{y}_n\end{bmatrix},\quad \boldsymbol{X}=\begin{bmatrix}x_{10}&x_{11}&\cdots&x_{1k}\\x_{20}&x_{21}&\cdots&x_{2k}\\\vdots&\vdots&\ddots&\vdots\\x_{n0}&x_{n1}&\cdots&x_{nk}\end{bmatrix},\quad \hat{\boldsymbol{\theta}}=\begin{bmatrix}\hat{\theta}_0\\\hat{\theta}_1\\\vdots\\\hat{\theta}_k\end{bmatrix},\quad \hat{\boldsymbol{\varepsilon}}=\begin{bmatrix}\hat{\varepsilon}_1\\\hat{\varepsilon}_2\\\vdots\\\hat{\varepsilon}_n\end{bmatrix}$$

Y 是实测值向量；$\hat{\boldsymbol{Y}}$是预测值向量；$\boldsymbol{X}$ 是输入变量矩阵，令 x_{10}，x_{20}，…，$x_{n0}=1$；$\hat{\boldsymbol{\theta}}$是参数向量；$\hat{\boldsymbol{\varepsilon}}$是残差向量，其每一个分量$\hat{\varepsilon}_i$ 都是预测值和实测值的差，也是真实误差 ε_i 的估计值。类似于一元线性回归方程，移项后可得

$$\hat{\boldsymbol{\varepsilon}}=\boldsymbol{Y}-\hat{\boldsymbol{Y}}=\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\theta}}$$

多元线性回归方程损失函数定义如下：

$$J(\hat{\boldsymbol{\theta}})=J(\hat{\theta}_0,\hat{\theta}_1,\cdots,\hat{\theta}_k)=\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\theta}})$$

$$=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-2\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\theta}}$$

依次计算 $J(\hat{\boldsymbol{\theta}})$对参数向量$\hat{\boldsymbol{\theta}}$的每一个分量$\hat{\theta}_j$ 的偏导数，并令偏导数等于 0，即可求得 $\hat{\theta}_j$，这些解组合成对应的参数向量$\hat{\boldsymbol{\theta}}$。

$$\frac{\partial}{\partial\hat{\boldsymbol{\theta}}}J(\hat{\boldsymbol{\theta}})=\frac{\partial(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-2\hat{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}}=-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\theta}}=0$$

多元线性回归方程参数向量$\hat{\boldsymbol{\theta}}$的正规方程解为

$$\hat{\boldsymbol{\theta}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

使用正规方程法求解参数向量$\hat{\boldsymbol{\theta}}$的必要条件是$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$的逆矩阵存在。

2. 梯度下降法

当k很大时，利用正规方程法计算矩阵的乘法和逆运算会变得很慢。另外，如果$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$是不可逆矩阵，不能直接使用正规方程法求解。这时可使用梯度下降法求解参数向量$\hat{\boldsymbol{\theta}}$。梯度下降法是一种寻找目标函数$J(\hat{\boldsymbol{\theta}})$最小化的迭代算法。先给定参数$\hat{\theta}_j$的初始值，通过不同的算法递推出新值，然后再以新值作为输入，继续迭代出一系列更新的值，如此反复由旧值递推出新值，直到使目标函数$J(\hat{\boldsymbol{\theta}})$极小的参数点处。

梯度下降法的具体步骤如下。

(1) 给$\hat{\theta}_j$一个随机的初始值。

(2) 计算损失函数$J(\hat{\boldsymbol{\theta}})$对参数$\hat{\boldsymbol{\theta}}$的每一个分量$\hat{\theta}_j$的偏导数，从而得到该分量在某点的梯度值。每做一次求偏导都表示找到了当前位置梯度最陡的方向，梯度的反方向就是函数值减小最快的方向。

(3) 结合$\hat{\theta}_j$的上一次值、步长与梯度递推出新值。递推公式如下：

$$\hat{\theta}_j=\hat{\theta}_j-\alpha\frac{\partial}{\partial\hat{\theta}_j}J(\hat{\boldsymbol{\theta}})\quad j=0,1,2,\cdots,k$$

求解上式的偏导，$\hat{\theta}_j$的迭代公式为

$$\hat{\theta}_j=\hat{\theta}_j-\alpha\sum_{i=1}^{n}(\hat{y}_i-y_i)x_{ij}\quad j=0,1,2,\cdots,k$$

其中，α是步长，也称学习率。在迭代过程中需要不断尝试不同的α，从而找到最合适的α。α值太小会导致迭代次数过多，$\hat{\boldsymbol{\theta}}$收敛过慢；$\alpha$值太大容易跳过$\hat{\boldsymbol{\theta}}$收敛值而在收敛值附近震荡。$x_{ij}$是第$i$个样本的第$j$个输入变量值，每一次迭代都需要把所有的$n$个样本都取到，把每一个样本的预测值减去实测值，再把这个差值乘上该样本的第j个特征值。

(4) 判断$\hat{\theta}_j$和损失函数$J(\hat{\boldsymbol{\theta}})$是否收敛。当两次迭代的值几乎不发生变化时，可判断收敛。如果没有收敛，重复步骤(2)和(3)，否则进行步骤(5)。

(5) 经过上述计算获得了其中一个参数的解，例如$\hat{\theta}_0$。然后重复以上步骤获得$\hat{\theta}_1$，$\hat{\theta}_2,\cdots,\hat{\theta}_k$，取得所有参数向量$\hat{\boldsymbol{\theta}}$的解，最终得到预测方程。

5.1.4 线性回归方程选择

1. 回归方程拟合优度

回归方程拟合优度是指回归直线对实测数据的拟合程度。如果全部实测点都落在回

归直线上，即可得到一个完美的拟合结果。但这种情况很少发生，大部分情况下，总有一些正的残差和一些负的残差围绕在回归直线周围。只需使围绕在回归直线的残差尽可能小，判定系数 R^2 就是这种拟合程度优劣的一个度量，通过对总离差平方和(Sum Squares of Total，SST)的分解得到判定系数 R^2。SST 定义如下：

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

SST 反映了实测值的总变化量，可分解为

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

残差平方和(Sum Squares Of Error，SSE)反映了由误差引起的实测值的变化。其定义如下：

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

若 SSE=0，则每个实测值都可由线性关系精确拟合，误差为 0。SSE 越大，实测值与预测值的偏差也越大，误差也越大。

模型平方和(Sum Squares of Model，SSM)反映了由输入变量引起的输出变量的变化。其定义如下：

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

若 SSM=0，则每个预测值均相等，即输出变量不随输入变量的变化而变化，二者之间的线性回归关系不成立，这实质上反映了 $\hat{\theta}_1=\hat{\theta}_2=\cdots=\hat{\theta}_k=0$。

因此，SST=SSM+SSE。SSM 越大，说明由线性回归关系描述输出变量总变化量的比例就越大，即输出变量与输入变量之间的线性关系越显著。

判定系数(determination coefficient)R^2 可以解释为输出变量实测值 $y_1,y_2,\cdots,y_n$ 的总变化量 SST 中被线性回归方程所描述的比例。其定义如下：

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

R^2 越大，说明线性回归方程描述输出变量总变化量的比例越大，从而 SSE 就越小，即拟合效果越好。

对于多元回归的情形，常用修正 R^2(Adj-R^2)代替 R^2，因为在模型中增加输入变量总能提高 R^2，Adj-R^2 考虑了加入模型的输入变量数，在比较不同多元模型时用 Adj-R^2 更合适。其定义为

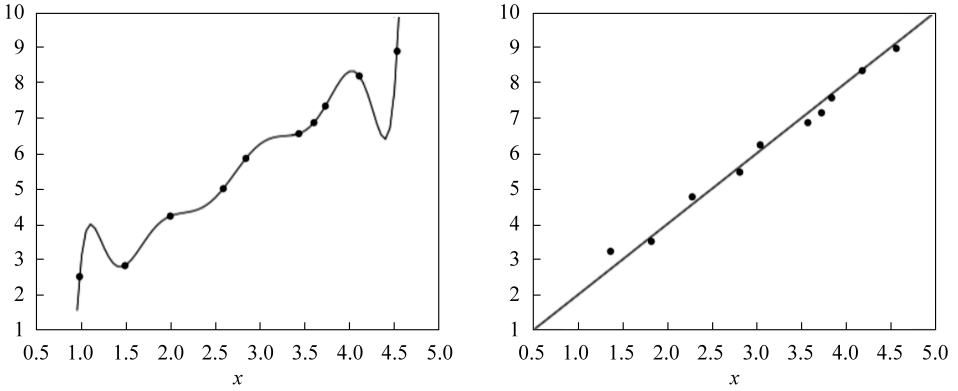
$$\text{Adj-}R^2=1-\frac{\mathrm{SSE}/(n-k-1)}{\mathrm{SST}/(n-f)}$$

若模型中包含 $\hat{\theta}_0$，则 $j=1$；否则 $j=0$。

2. 防止过拟合

因样本数据不可避免地存在测量误差和噪声数据，若一味地追求高拟合优度，很可能

导致过拟合。例如，图 5.3 中的 10 个点可用多项式$\hat{y}=\hat{\theta}_0+\hat{\theta}_1x+\hat{\theta}_2x^2+\cdots+\hat{\theta}_9x^9$ 拟合，也可用线性回归模型$\hat{y}=2x$ 拟合。显然图 5.3(a)拟合了所有噪声数据，出现了过拟合现象。从图形上看，是一条曲线且曲线几乎通过所有的实测数据点；从表达式上看，是一个高次函数，输入变量系数值波动较为明显，过拟合模型在未知数据上的预测效果很差，模型泛化能力低。



(a) 过拟合　　(b) 正常拟合

图 5.3　过拟合与正常拟合

此外，当模型的输入变量过多且样本数据过少时也会出现过拟合问题，这种情况下一般需要采用正则化处理。正则化处理采用某种约束为参数规定一个取值范围，减小输入变量的数量级，防止参数出现较大波动，从而在尽可能符合数据原始分布的基础上得到一个平滑、简单的模型，达到减化模型、降低过拟合可能性的目的。

正则化处理的表现形式是在线性回归损失函数 $J(\hat{\boldsymbol{\theta}})$中加正则化项。正则化一般有 L1 正则化与 L2 正则化。

L1 正则化在线性回归中称为 Lasso 回归，即常说的套索回归。其方法是：在损失函数 $J(\hat{\boldsymbol{\theta}})$后增加参数向量$\hat{\boldsymbol{\theta}}$的 L1 范数，即所有输入变量系数绝对值的和来实现正则化。函数表达式如下：

$$J(\hat{\boldsymbol{\theta}})=\frac{1}{2n}\left[\sum_{i=1}^{n}(\hat{y}_i-y_i)^2+\lambda\sum_{j=1}^{k}|\hat{\theta}_j|\right]$$

其中，n 是样本个数；k 是输入变量的个数；λ 是正则化参数，用来调节损失函数的误差项和正则化项的权重。L1 正则化可以使一些输入变量的系数变小，甚至还使一些绝对值较小的系数直接变为 0，以增强模型的泛化能力。

L2 正则化在线性回归中也称为 Ridge 回归，即常说的岭回归。其方法是：在损失函数 $J(\hat{\boldsymbol{\theta}})$后增加参数向量$\hat{\boldsymbol{\theta}}$的 L2 范数的平方项，即所有输入变量系数的平方和来实现正则化。函数表达式如下：

$$J(\hat{\boldsymbol{\theta}})=\frac{1}{2n}\left[\sum_{i=1}^{n}(\hat{y}_i-y_i)^2+\lambda\sum_{j=1}^{k}\hat{\theta}_j^2\right]$$

L2 正则化会对输入变量系数按比例进行缩放，而不是像 L1 正则化那样减去一个固定值，这使得输入变量系数变小而不会变为 0，因此 L2 正则化会让模型变得更简单，防止过拟合，而不会起到特征选择的作用。

损失函数 $J(\hat{\boldsymbol{\theta}})$经过了正则化处理，可使用梯度下降法和正规方程法求解参数向量 $\hat{\boldsymbol{\theta}}$。以 L2 正则化函数为例，参数向量$\hat{\boldsymbol{\theta}}$的解如下：

(1) 正规方程法：

$$\hat{\boldsymbol{\theta}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda\begin{bmatrix}0&&&\\&1&&\\&&1&\\&&&\ddots\\&&&&1\end{bmatrix}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

(2) 梯度下降法：

$$\hat{\theta}_0=\hat{\theta}_0-\alpha\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)x_{i0}$$

$$\hat{\theta}_j=\hat{\theta}_j-\alpha\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)x_{ij}+\frac{\lambda}{n}\hat{\theta}_j\right]\quad j=0,1,2,\cdots,k$$

3. 回归方程总体显著性检验

在实际问题的研究中，人们事先可能并不确定输出变量 $\boldsymbol{Y}$ 与输入变量 $x_1,x_2,\cdots,x_k$ 之间的关系。在求解参数向量$\hat{\boldsymbol{\theta}}$之前，用线性回归方程拟合某个样本数据可能效果较好，但是对总体数据的效果却未必同样好，通过样本估计的关系也不一定能延伸到总体关系。因此，在求解线性回归方程后，还需要对回归方程进行总体的显著性检验，即从整体上看输入变量 $x_1,x_2,\cdots,x_k$ 对输出变量 $\boldsymbol{Y}$ 是否有显著影响，为此提出如下假设：

原假设 H_0：$\theta_1=\theta_2=\cdots=\theta_k=0$。

备择假设 H_1：$\theta_1,\theta_2,\cdots,\theta_k$ 不全为 0。

检验统计量：$F=\dfrac{\mathrm{SSM}/k}{\mathrm{SSE}/(n-k-1)}$。

当原假设 H_0 成立时，检验统计量服从自由度为$(k,n-k-1)$的 F 分布，对于给定的显著性水平 α，当统计量的观测值 F_0 大于临界值 $F_c(k,n-k-1)$时，拒绝原假设 H_0，即，在显著性水平 α 下，输出变量与输入变量之间的线性回归关系是显著的，或称回归方程是显著的；否则不能拒绝 H_0。也可以根据 $p=P\{F\geqslant F_0\}$值得出结论：若 p 值小于给定的显著性水平 α，则拒绝原假设 H_0。

4. 回归方程系数显著性检验

即使回归方程通过显著性检验，也只能表明参数向量 θ 的各个分量不全为零，并不意味着每个输入变量 $x_1,x_2,\cdots,x_k$ 对输出变量 $\boldsymbol{Y}$ 的影响都显著，因此需要对每个输入变量

$x_1,x_2,\cdots,x_k$ 进行显著性检验。若某个系数 $\theta_j=0$，则 x_j 对 $\boldsymbol{Y}$ 的影响不显著。因此，下面考虑从回归方程中剔除 X_j。

原假设 $H_0^{(j)}:\theta_j=0$。

备择假设 $H_1^{(j)}:\theta_j\neq 0,j=1,2,\cdots,k$。

检验统计量：$t_j=\dfrac{\hat{\theta}_j}{\sqrt{\mathrm{SSE}/(n-k-1)}}$。

当原假设 $H_0^{(j)}$ 成立时，检验统计量服从自由度为$(n-k-1)$的 t 分布。对于给定的显著性水平 α，当 t_j 统计量的观测值 t_{j0} 大于临界值 $t_{jc}(n-k-1)$时，拒绝原假设 $H_0^{(j)}$，即，在显著性水平 α 下，θ_j 不为 0，认为 x_j 对 $\boldsymbol{Y}$ 的作用是显著的；否则不能拒绝 θ_j 为 0。也可以根据 $p_j=P\{|t_j|\geqslant|t_{j0}|\}$值得出结论：若 p_j 值小于给定的显著性水平 α，则拒绝原假设 H_0。

5. 回归诊断

残差分析和共线性诊断是对回归模型进行回归诊断的两种常用方法。残差分析的目的是检验线性回归方程的可行性，包括残差项的等方差、独立性和正态性的假设。共线性诊断用来确定多元回归中多个输入变量之间是否存在多重共线性。

1）残差分析

通过绘制残差图可判定残差项的等方差和独立性。以残差为纵坐标，以实测值 y_i、预测值 $\hat{y}_i$、输入变量 $x_j(j=1,2,\cdots,k)$或序号、观测时间等为横坐标绘制的散点图称为残差图。正常残差图中的散点应围绕 0 随机波动且变化幅度在一条矩形带内。图 5.4 给出了 3 个残差图。左图中对于所有横坐标轴变量的值，残差的方差都相同，描述输出变量和输入变量之间关系的回归模型是合理的；中图表明残差的方差随变量的增大而增大，违背了残差等方差的假设；右图表明所选的回归模型不合理，应包含输入变量的二次项。

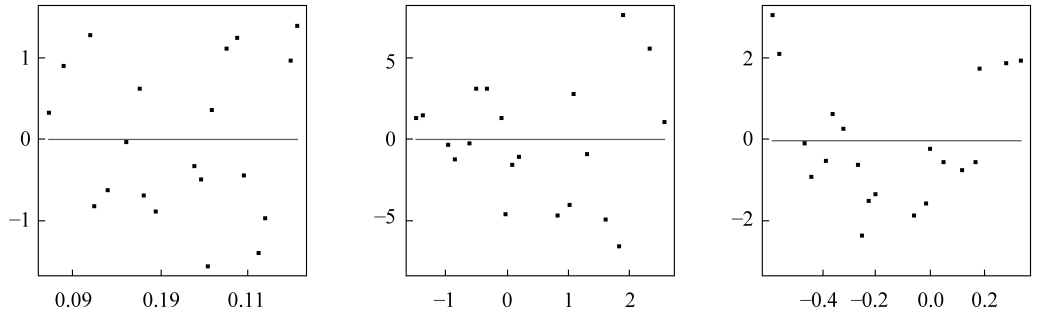


图 5.4 残差图示例

2）共线性诊断及解决方法

如果模型中包含的输入变量过多，输入变量之间就可能存在近似线性关系，这种现象称为输入变量间的多重共线性，简称共线性。下列情况表明可能存在共线性问题：

（1）回归方程的 F 检验通过，而部分回归系数 θ_j 的检验未通过。

(2) 回归系数的正负号与预期的相反。

(3) 模型中增加或删除一个输入变量对输入变量系数的估计值影响显著。

研究结果表明,产生这些问题的原因之一就是输入变量之间存在共线性,检测共线性严重程度的一种方法是计算矩阵的条件数(condition number)。条件数用来度量线性模型(或者矩阵)的稳定性或者敏感度。如果一个线性模型的条件数小于 20,就属于比较稳定的;否则就是不稳定的,其输出结果可信度不高。

如果发现线性模型中存在共线性问题,可进行手动筛选,还可以通过算法自动进行输入变量的筛选,常用的算法包括向前法、向后法和逐步选择法。逐步选择法是交互地引入和剔除输入变量,每次引入对输出变量影响最显著的输入变量后,都对模型中的已有输入变量重新进行显著性检验,把由于新输入变量的引入而变得对输出变量影响不显著的输入变量剔除,然后再考虑下一轮输入变量的引入和剔除,直到既不能引入也不能剔除输入变量为止。输入变量的每一次引入和剔除其实都是在做假设检验,在实际应用中也可以直接通过输入变量各项回归系数的正负和大小判断输入变量与输出变量的相关性,但通常还是应用逐步选择的方法。

5.1.5 线性回归方程预测

当通过前述多种检验证明一个回归方程的线性关系显著,即拟合效果较好时,便可利用最优线性回归方程根据输入变量的取值估计或预测输出变量的取值。预测或估计的类型主要有两种:点估计和区间估计。

1. 点估计

点估计是指对输出变量求点估计值。常用的点估计方法有两种,即平均值的点估计和个体值的点估计,二者使用的公式相同,区别在于表述的意义不同。

将输入变量的一组新实测值 $\boldsymbol{x}_0=(x_{01},x_{02},\cdots,x_{0k})$ 代入回归方程,对应的输出变量的预测值为 $E(\boldsymbol{y}_0)=\hat{\boldsymbol{y}}_0=\hat{\theta}_0+\hat{\theta}_1x_{01}+\cdots+\hat{\theta}_jx_{0k}$。如果把该值理解为输出变量平均值的一个点估计值,则该值是一个期望值;如果把该值理解为输出变量个体值的一个点估计值,则该值是一个具体值。

2. 区间估计

区间估计是指利用最优线性回归方程,对输入变量的特定值 $\boldsymbol{x}_0=(x_{01},x_{02},\cdots,x_{0k})$,求出输出变量的一个估计值的区间。与点估计类似,区间估计分为两种。

(1) 输出变量平均值的置信区间估计:

$$(\hat{\boldsymbol{y}}_0-t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{x}_0(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{x}_0^{\mathrm{T}}},\hat{\boldsymbol{y}}_0+t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{x}_0(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{x}_0^{\mathrm{T}}})$$

其中,$s=\sqrt{\mathrm{MSE}}$,n 为观测次数,k 为输入变量个数。

(2) 输出变量个体值的预测区间估计:

$$(\hat{\boldsymbol{y}}_0-t_{\alpha/2}(n-k-1)s\sqrt{1+\boldsymbol{x}_0(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{x}_0^{\mathrm{T}}},\hat{\boldsymbol{y}}_0+t_{\alpha/2}(n-k-1)s\sqrt{1+\boldsymbol{x}_0(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{x}_0^{\mathrm{T}}})$$