第七章 回归分析

作为现代统计学的重要分支，回归分析（Regression Analysis）是人们对自然世界的数据背后规律的最早探索。人们通过搜集数据，将想要分析的变量放在一起建立模型，实现对关注结果的预测与分析，如对未来钢铁的产量增长率以及一些难以实时测量的质量指标、成分参数等进行预测。

本章首先介绍回归分析的基本概念，再着重介绍经典的线性回归方法（最小二乘估计），并针对多重共线性问题展开讨论，介绍岭回归、LASSO 回归，主元回归（Principal Component Regression, PCR）和偏最小二乘回归（Partial Least Squares Regression, PLSR），以及一些非线性回归方法，最后通过对模型验证进行介绍，以保证回归模型的有效性。

7.1 回归分析的基本概念

回归分析是广泛用于分析多因子数据的方法之一。回归分析使用方程来表达所感兴趣的变量（响应变量）与一系列相关预测变量之间的关系，本小节将介绍回归分析相关的基础概念以及流程。

7.1.1 导引

回归分析是研究变量间函数关系的一种方法，变量之间的这种关系可以表示为方程或模型的形式，该方程或模型将响应变量与一个或多个预测变量联系起来。预测变量也可以称为解释变量、独立变量、协变量、回归变量或因素。虽然我们经常使用独立变量这种名称，但由于实际中的预测变量之间很少是相互独立的，所以这个名称并不贴切。本章节后续均称为预测变量。

我们用 Y 表示响应变量，用X_1，X_2，…，X_p表示预测变量，其中p是预测变量的个数，Y 和X_1，X_2，…，X_p之间的真实关系可近似地用下述回归模型刻画

$$Y = f(X_1，X_2，\cdots，X_p) + \varepsilon，\tag{7.1}$$

其中ε是随机误差，它代表在近似过程中产生的偏差，也就是模型不能精确拟合数据的原因。函数$f(X_1，X_2，\cdots，X_p)$刻画了 Y 和X_1，X_2，…，X_p之间的关系，最简单的情形是线性回归模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon，\tag{7.2}$$

其中β_1，$\beta_2 \cdots$，β_p称为回归系数，β_0称为截距，它们都是未知常数，称为模型的回归参数，这些未知参数可由数据估计。

使用回归模型的目的有很多，最常见的有以下几种：

1）回归分析可以建立模型来描述数据集。比如，通过大量工业生产时间和产品生产量的数据得到的回归模型，相比数据表甚至数据图形，都更加方便和实用。

2）回归方法可以解决参数估计问题。举例来说，化学工程中使用米氏方程$y=\frac{\beta_1 x}{x+\beta_2}+\varepsilon$来描述反应速率$y$与浓度$x$的关系。在这一模型中，$\beta_1$是反应的最终速率，即随着浓度的增大即能达到的最大值。如果得到了由不同浓度下速率的观测值组成的样本，那么化学工程设计中就能通过回归分析来得到能拟合数据的模型，从而得到最大速率的估计值。

3）回归的很多应用领域都涉及对响应变量的预测。比如，我们可能希望预测生产一定数量的工业产品所需的生产时间，这一预测可能有助于规划生产活动，如设计出货数量、安排未来生产计划，也可能有助于评估生产作业的效率。

4）使用回归模型也可以进行控制。比如，化学工程中使用回归分析来得出有关纸张抗张强度与木浆中硬木浆浓度的模型，然后利用这一方程，通过改变硬木浆的浓度水平，控制抗张强度使其达到合适的值。以控制为目的使用回归方程时，重要的是变量之间要存在因果关系。注意，如果仅使用方程进行预测，因果关系可能并不是必要的。但用于构建回归方程的原始数据中存在的因果关系是必要的。

举例来说，在高炉煤气调节中，需要使用煤气利用率历史数据用以建模预测未来数据，从而对现有炉况状态进行相应的调整。因此，需要使用历史高炉操作参数进行相应的拟合与预测，从而对未来炉况的发展情况做相应判断，然后通过改变相应的高炉操作使高炉保持在一个稳定的炉况范围内。针对此种情况，则需要建立高炉煤气利用率回归模型。

7.1.2 回归建模的分类及步骤

回归分析包括以下步骤：问题陈述、选择变量、收集数据、模型设定、选择拟合方法、模型拟合、模型的评估与选择等，下面详细介绍这些步骤。

（1）问题陈述

回归分析通常是从对问题的陈述开始的，也就是要确定需要分析研究哪些问题。如果把精力浪费在一个陈述模糊或陈述错误的问题上，就会导致选择错误的变量集或统计分析方法，也会导致选择错误的模型。如果想计算未来高炉煤气利用率，就应该考虑影响该利用率的相关变量的历史数据。如果其他变量未发生波动，某个相关变量增大会直接导致高炉煤气利用率波动，则应该把该变量设为预测变量，高炉煤气利用率作为响应变量。

（2）选择相关变量

问题陈述清楚之后，可根据该研究领域专业人士的意见或者关联分析和因果分析等数据分析的方法选择变量集合，获得所有可以用来解释或预测响应变量的预测变量。例如，在高炉煤气调节中，冷风流量、冷风压力、热风压力、富氧流量、富氧压力、喷煤量、边缘矿焦比 3、边缘矿焦比 4、中心焦比 7、中心焦比 11 等变量，对高炉煤气利用率有着较大的关系，则选择这些操作变量作为相关变量。

（3）收集数据

选择好潜在的相关变量后，下一步是从实际中收集分析问题使用的数据。在每种情况下，我们收集到n个目标的观测数据。每个目标的观测数据都是对该目标所有潜在的相关变量的测量值。

收集到的数据通常如表 7-1 进行记录。表 7-1 的每一列代表一个变量，而每一行表示一个观测，对应某个目标的$p+1$个值，其中一个为响应变量的值y_i，其他p个预测变量中的每一个对应一个值。符号x_{ij}指第j个预测变量的第i个观测值，即第一个下标对应观测序号，第二个下标对应预测变量的序号。表 7-1 中的每个变量按其取值情况可以分为定量变量或定性变量。

表 7-1 回归分析中数据的变量符号

观测序号	响应变量	预测变量			
	Y	X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

（4）模型设定

为了将响应变量和预测变量联系起来，通常先从文献或者由专家根据他们的知识或主客观判断得到当前模型的形式。要注意的是，此处只需给出模型的形式，它可以含有未知参数。选择式（7.1）中函数$f(X_1，X_2，\cdots，X_p)$的形式。该函数可以分为两类：线性和非线性。

线性函数如

$$Y=\beta_0+\beta_1X_1+\varepsilon \tag{7.3}$$

非线性函数如

$$Y=\beta_0+e^{\beta_1X_1}+\varepsilon \tag{7.4}$$

注意这里的线性项（或非线性项）不是表示Y与X_1，X_2，…，X_p之间的关系，而是指等式关于回归参数是线性的（或非线性的）。

只有一个响应变量的回归分析称为单变量回归，有两个或两个以上响应变量的回归称为多变量回归。简单回归和多元回归并不是单变量回归和多变量回归。他们的区别在于：（1）简单回归和多元回归是由预测变量的个数决定的：简单回归只有一个预测变量，多元回归有两个或两个以上预测变量；（2）单变量回归和多变量回归是由响应变量的个数决定的：单变量回归只有一个响应变量，多变量回归有两个或两个以上响应变量。

除此之外，还有其他回归分析类型，如线性回归与非线性回归是由响应变量和预测变量之间是否具有非线性关系而区分的。如果所有预测变量都是定性变量，分析这些数据的方法称为方差分析。如果预测变量有定量变量也有定性变量，此时的回归分析称为协方差分析。各种回归分析的分类详情见表 7-2。

表 7-2 回归分析类型

回归类型	条件
单变量回归	只有一个定量的响应变量
多变量回归	有两个或两个以上定量的响应变量
简单回归	只有一个预测变量
多元回归	有两个或两个以上预测变量
线性回归	方程关于所有的参数都是线性的，或经变量变换后是线性的
非线性回归	响应变量和某些预测变量之间具有非线性关系，或一些参数是以非线性形式出现且不能经变换将参数线性化
方差分析	预测变量都是定性变量
协方差分析	预测变量有定量变量，也有定性变量
逻辑回归	响应变量是定性变量

（5）拟合方法选择

确定模型和收集数据之后，接下来是利用数据估计模型参数，也称为参数估计或模型拟合。最常用的估计方法是最小二乘法，该方法在某些假设下有很多好的性质。本章节中主要采用最小二乘法和它的一些变形方法，如加权最小二乘法。在某些情况下，例如当一个或多个假设不成立时，其他估计方法可能会优于最小二乘法。本章节考虑的其他估计方法还有极大似然估计法、岭回归法以及主成分回归法等。

（6）模型拟合

利用选定的估算方法（最小二乘法）和收集到的数据进行回归参数估计或

模型拟合。式（7.2）中回归参数β_0，β_1，…，β_p的估计用$\hat{\beta}_0$，$\hat{\beta}_1$，…，$\hat{\beta}_p$表示。于是，回归方程的估计可以写成

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots \hat{\beta}_p X_p \tag{7.5}$$

参数上方的记号“^”表示该参数的估计，$\hat{Y}$为拟合值。注意式（7.5）还可以用预测变量的任意值来预测相应响应变量的值。这种情况下获得的Y称为预测值。拟合值和预测值的不同在于，拟合值对应预测变量的值就是数据中的某个观测，而预测值对应的可以是预测变量的任何取值。

（7）模型评价

模型的有效性依赖于某些假设，通常是指对数据和模型的假设，对分析和结论的准确性至关重要的是这些假设条件是否满足。例如，在用（7.5）做任何分析之前，我们首先需要确定特定的假设是否成立。我们需要解决以下问题：

① 需要哪些假设？

② 对于每个假设，我们如何确定该假设是否满足？

③ 当一个或更多假设不成立时，我们该如何处理？

本章节的后续内容将详细介绍标准的回归假设以及回答上面的问题。本章节后续将从线性回归模型入手，主要介绍最小二乘法，再针对最小二乘法的不足，介绍其补充方法。

7.2 线性回归

线性回归就是研究响应变量Y和预测变量X之间的关系，首先使用协方差和相关系数来刻画变量间线性关系的方向和强度，然后建立线性回归模型。

7.2.1 线性回归模型

假定响应变量Y和预测变量X之间的关系可用如下的线性模型刻画

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{7.6}$$

其中是β_0常数项，β_1是模型回归系数，它们都是常数，称为模型参数，ε是随机扰动或误差。假设误差项的均值为 0，且方差σ^2未知。

线性方程（7.6）是对Y和X之间真实关系的一种近似，即Y与X的关系可以用一个X的线性函数近似地表示，ε是这种近似的偏差。要特别说明的是，ε只是随机误差，不包含Y与X之间关系的任何信息。

参数β_0和β_1通常称为回归参数，斜率β_1是由X的变化一单位所产生的Y均值分布的变化率。如果数据中X的范围包括$X = 0$，那么截距β_0是$X = 0$时响应变量Y均值的分布；如果X的范围不包括 0，那么β_0没有实际含义。

包含多于一个回归变量的回归模型称为多元回归模型，多元线性回归的结

果是对简单线性回归结果的拓展。一般情况下，响应变量y可以与p个回归变量即预测变量相关，其模型为

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

称为有p个回归变量的多元线性回归模型。参数$\beta_j (j = 0,1,\cdots,p)$称为回归系数。这一模型描述了回归变量$x_j$组成的$p$维空间中的一个超平面。参数表示当其他回归变量$x_i (i \neq j)$都保持不变时，$x_j$每变化一单位值，响应变量$y$均值的变化的期望，通常$\beta_j (j = 0,1,\cdots,k)$也称为偏回归系数。回归方程拟合即回归模型拟合，一般用于预测响应变量y的未来观测值或估计响应变量y在特定水平下的均值。

7.2.2 最小二乘估计

在某些假设下，最小二乘估计有很多好的性质。故本节采用最小二乘估计法来完成模型参数的估计。

当进行假设检验或构造置信区间时，必须假设由$x_1, x_2, \cdots, x_p$给定的y的条件分布是正态分布，其中均值为$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}$，方差为$\sigma^2$。

则将对应的样本回归模型写为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_{ij} + \varepsilon_i$$

最小二乘函数为

$$S(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2$$

必须对函数S做关于$\beta_0, \beta_1, \cdots, \beta_p$的最小化。

得到最小二乘正规方程为

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7.7}$$

式中：

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{7.8}$$

其中，正规方程的解将是最小二乘估计量 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$。使用矩阵记号表达正规

方程对处理多元回归模型更加方便。

求最小二乘法估计量向量$\widehat{\boldsymbol{\beta}}$，最小化值为

$$S(\boldsymbol{\beta})=\sum_{i=1}^{n}\varepsilon_i^2=\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}=(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$$

最小二乘估计量必须满足

$$\frac{\partial S}{\partial\boldsymbol{\beta}}\Big|_{\widehat{\boldsymbol{\beta}}}=-2\boldsymbol{X}'\boldsymbol{y}+2\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}=0 \tag{7.9}$$

简化为

$$\boldsymbol{X}'\boldsymbol{y}=\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} \tag{7.10}$$

该方程为最小二乘正规方程，因此$\widehat{\boldsymbol{\beta}}$估计值为

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{7.11}$$

最终可写出拟合的最小二乘回归方程为

$$\hat{\boldsymbol{Y}}=\hat{\beta}_0+\hat{\beta}_1\boldsymbol{X_1}+\cdots+\hat{\beta}_p\boldsymbol{X_p} \tag{7.12}$$

对于预测变量的每一组观测，我们可以计算y_i的拟合值

$$\hat{y_i}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}+\varepsilon_1 \tag{7.13}$$

$\boldsymbol{\varepsilon}$的方差σ^2的无偏估计为

$$\hat{\sigma}^2=\frac{S(\boldsymbol{\beta})}{n-p-1} \tag{7.14}$$

$S(\beta)$也称为残差平方和（Sum of Squared Error，SSE），分母中的$n-p-1$是自由度（df）, 等于观测数据个数减去待估参数的个数。

在某些标准的回归假设下，最小二乘估计具有下面的性质：

1）$\hat{\beta}_j(j=0,1,\cdots,k)$是$\beta_j$的无偏估计，并且方差为$\sigma^2c_{jj}$，其中$c_{jj}$是正规方程组的系数矩阵（称为平方和及乘积和矩阵）的逆矩阵 C 的主对角线上的第j个元素。在$\boldsymbol{\beta}$所有的线性无偏估计中，最小二乘估计的方差是最小的。所以，将最小二乘估计称为最佳线性无偏估计（Best Linear Un- biased Estimators, BLUE）。

2）$\hat{\beta}_j,j=0,1,\cdots,k$服从均值为$\beta_j$，方差为$\sigma^2c_{jj}$的正态分布。

3）$W=\mathrm{SSE}/\sigma^2$服从自由度为$n-p-1$的χ^2分布，并且每个$\hat{\beta}_j,j=0,1,\cdots,k$和$\sigma^2$相互独立。

前面拟合的回归方程可以用来做预测。首先，我们先区分两种类型的预测：

1）对于任意给定的预测变量的值$\boldsymbol{X_0}=(x_{01},x_{02},\cdots,x_{0p})$，给出响应变量$Y$的预测值。

2）当$\boldsymbol{X}=\boldsymbol{X_0}$时，估计响应变量的响应均值$\mu_0$。

对于第一个问题，预测值$\hat{y}_0$是

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\hat{\beta}_2x_{02}+\cdots+\hat{\beta}_px_{0p} \tag{7.15}$$

对于第二个问题，实际上是响应均值μ_0的估计问题，由下式估计：

$$\hat{\mu}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\hat{\beta}_2x_{02}+\cdots+\hat{\beta}_px_{0p} \tag{7.16}$$

比较式（7.15）和式（7.16）可以发现，μ_0的点估计和响应变量的预测值$\hat{y}_0$是一样的。而比较其标准误可以发现，$\hat{\mu}_0$的标准误小于$\hat{y}_0$的标准误。当$\boldsymbol{X}=\boldsymbol{X_0}$时，预测一个观测值比估计响应均值存在更大的不确定性。估计响应均值就是一种取平均的做法，通过平均来减少波动性和不确定性。

例 7.1 通过相关性分析，选取富氧压力和热风压力作为预测变量，冷风压力作为响应变量，收集其中 1000 组数据。用多元线性模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$$

拟合数据。

散点图在拟合多元回归模型中是很有用的。图 7.1 为点图矩阵，为二位散点图的二维阵列，其中（除对角线外）每个图框包含一个散点图。从每个散点图可以观察到富氧压力、热风压力与冷风压力两两之间成线性的关系。

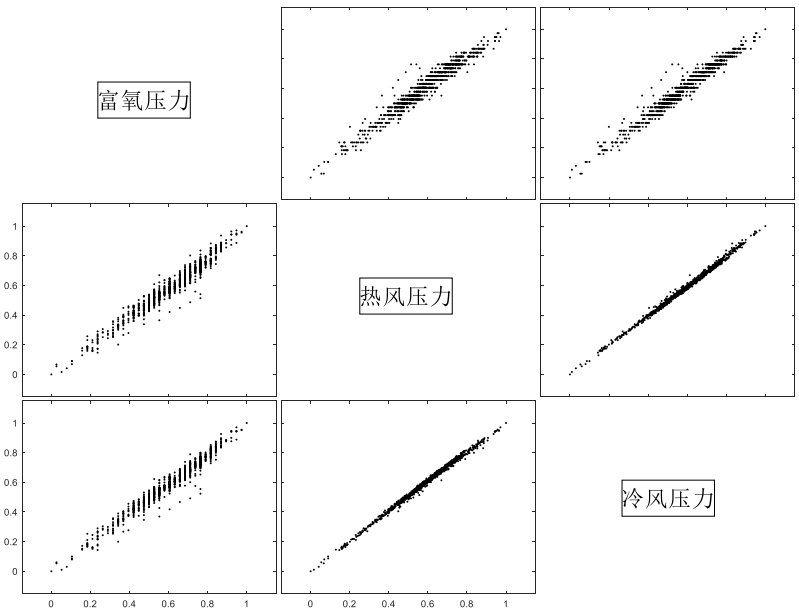


图 7.1 散点图矩阵

当只有两个回归变量时，三维散点图有时对响应变量与回归变量之间关系的可视化是有用的。图 7.2 显示的三维散点图表明该多元线性回归模型可以为

以富氧压力和热风压力作为预测变量，以冷风压力作为响应变量提供合理的数据拟合。

为了拟合多元回归模型，先形成矩阵$\boldsymbol{X}$与向量$\boldsymbol{y}$分别为

$$\boldsymbol{X}=\begin{bmatrix}\mathbf{1} & x_{11} & x_{21}\\ \vdots & \vdots & \vdots\\ \mathbf{1} & x_{11000} & x_{21000}\end{bmatrix}$$

$$\boldsymbol{y}=\begin{bmatrix}y_{\mathbf{1}}\\ \vdots\\ y_{\mathbf{1000}}\end{bmatrix}$$

$\boldsymbol{\beta}$的最小二乘估计量为

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

计算易得

$$\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}=\begin{bmatrix}0.0024\\ 0.1543\\ 0.8491\end{bmatrix}$$

最小二乘拟合为

$$\hat{y}=0.0024+0.1543x_1+0.8491x_2$$

当前拟合值的总体均方根误差为 RMSE=0.0513。证明当前多元线性回归模型的拟合效果较为优秀，即富氧压力和热风压力作为预测变量，可以完成对冷风压力的预测。

以富氧压力为例，说明在其他因素不变的情况下，富氧压力每增加一个单元值，冷风压力下降 0.15 个单位值。

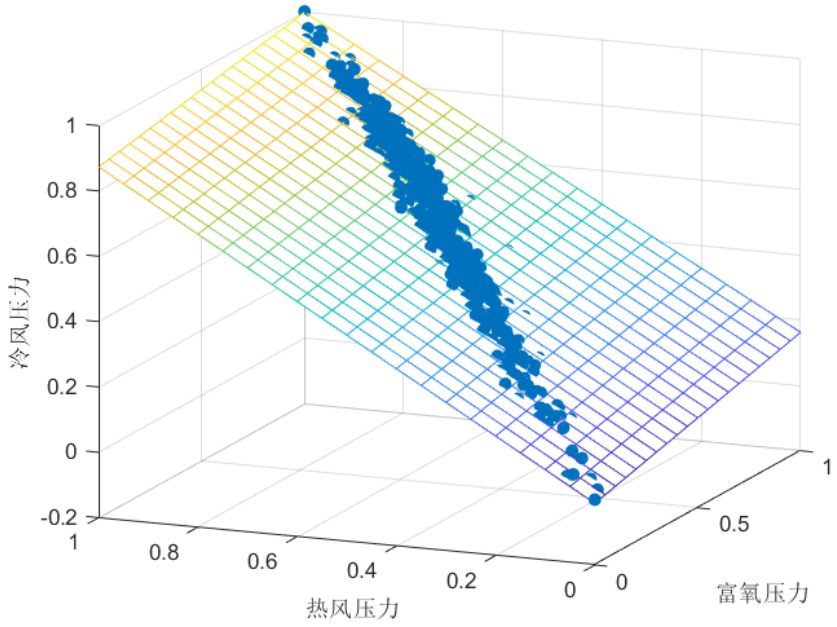


图 7.2 三维散点图

7.2.3 加权最小二乘估计

在违背某些内含假设的情况下，加权最小二乘法在构建回归模型时也是有用的。本节将解释当原始最小二乘不满足方差相等的假设时，如何使用加权最小二乘。

当误差$\boldsymbol{\varepsilon}$不相关但方差不相等，$\boldsymbol{\varepsilon}$的协方差矩阵为

$$\sigma^2\boldsymbol{V}=\sigma^2\begin{bmatrix}\frac{1}{w_1} & & & 0\\ & \frac{1}{w_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{w_n}\end{bmatrix} \tag{7.17}$$

其估计通常称为加权最小二乘。令$\boldsymbol{W}=\boldsymbol{V}^{-1}$。由于$\boldsymbol{V}$是对角矩阵，所以$\boldsymbol{W}$也是对角的，其对角线元素即权重为$w_1,w_2,\cdots,w_n$。由最小二乘正规方程可得，加权最小二乘正规方程为

$$(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})\widehat{\boldsymbol{\beta}}=\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} \tag{7.18}$$

这是多元回归的加权最小二乘正规方程，类似于一元回归。因此，

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} \tag{7.19}$$

为加权最小二乘估计量。注意方差较大的观测值将比方差较小的观测值有更小的权重。

加权最小二乘估计量可以简单地通过普通最小二乘的计算机程序得到。如果将第i次观测的每个观测值（为了方便解释将 1 包括在内）乘以该观测值权重的平方根，那么会得到变换后的数据集

$$\boldsymbol{B}=\begin{bmatrix}1\sqrt{w_1} & x_{11}\sqrt{w_1} & \cdots & x_{1k}\sqrt{w_1}\\ 1\sqrt{w_2} & x_{21}\sqrt{w_2} & \cdots & x_{2k}\sqrt{w_2}\\ \vdots & \vdots & & \vdots\\ 1\sqrt{w_n} & x_{n1}\sqrt{w_n} & \cdots & x_{nk}\sqrt{w_n}\end{bmatrix},\begin{bmatrix}y_1\sqrt{w_1}\\ y_2\sqrt{w_2}\\ \vdots\\ y_n\sqrt{w_n}\end{bmatrix} \tag{7.20}$$

现在如果对这一变换后的数据做普通最小二乘，那么所得到的

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{z}=(\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} \tag{7.21}$$

为$\boldsymbol{\beta}$的加权最小二乘估计量。

总的来说，加权最小二乘法是对原模型进行加权，使之成为一个新的不存在异方差性的模型，然后采用普通最小二乘法估计其参数的一种数学优化技术。

7.2.4 极大似然估计

最小二乘法可以用于线性回归模型的参数估计，产生最佳线性无偏估计量，此时不对误差ε的分布形式做任何假设。但对于其他统计过程，比如假设检验与置信区间构造，都假设误差服从正态分布。如果误差的分布形式已知，那么

就可以使用另一种参数估计方法--极大似然法。

考虑数据(y_i, x_i)，$i = 1,2,\cdots,n$。假设回归模型中的误差服从正态独立分布NID$(0,\sigma^2)$，那么样本的观测值y_i服从均值为$\beta_0 + \beta_1 x_i$，方差为σ^2的正态分布且独立。似然函数由观测值的联合分布得到。如果考虑给定观测值的联合分布，参数β_0、β_1及σ^2为未知参数，那么就有极大函数对于误差服从正态分布的简单线性回归模型而言，其似然函数为

$$\begin{aligned} L(y_i, x_i, \beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp[-\tfrac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2] \\ &= (2\pi\sigma^2)^{-1/2} \exp[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2] \end{aligned} \tag{7.22}$$

极大似然估计量的参数值记为 $\tilde{\beta}_0$、$\tilde{\beta}_1$ 与$\tilde{\sigma}^2$。最大化L或与其等价的$ln\, L$为

$$\begin{aligned} ln\, L\,(y_i, x_i, \beta_0, \beta_1, \sigma^2) &= -\left(\frac{n}{2}\right) ln\, 2\,\pi - \left(\frac{n}{2}\right) ln\, \sigma^2 \\ &= -\left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned} \tag{7.23}$$

而极大似然估计量 $\tilde{\beta}_0$、$\tilde{\beta}_1$ 与$\tilde{\sigma}^2$必须满足

$$\begin{aligned} \left.\frac{\partial\, ln\, L}{\partial \beta_0}\right|_{\tilde{\beta}_0 \cdot \tilde{\beta}_1 \cdot \tilde{\sigma}^2} &= \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right) = 0 \\ \left.\frac{\partial\, ln\, L}{\partial \beta_1}\right|_{\tilde{\beta}_0 \cdot \tilde{\beta}_1 \cdot \tilde{\sigma}^2} &= \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right) x_i = 0 \end{aligned} \tag{7.24}$$

以及

$$\left.\frac{\partial\, ln\, L}{\partial \sigma^2}\right|_{\tilde{\beta}_0 \cdot \tilde{\beta}_1 \cdot \tilde{\sigma}^2} = -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\sum_{i=1}^{n}\left(y_1 - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)^2 = 0 \tag{7.25}$$

方程给出了极大似然估计量为

$$\tilde{\beta}_0 = \overline{y} - \tilde{\beta}_1 \overline{x} \tag{7.26}$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{7.27}$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n} \tag{7.28}$$

正如简单线性回归的情形，可以证明当模型误差为独立正态分布时，多元线性回归中模型参数的最大似然估计量也是其最小二乘估计量。模型为

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{7.29}$$

且误差为独立正态分布，其方差为常数σ^2，即ε服从$N(\mathbf{0}, \sigma^2 \boldsymbol{I})$。误差的正态密度函数为

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right) \tag{7.30}$$

最大似然函数是 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$或$\prod_{i=1}^{n} f\,(\varepsilon_i)$的联合密度函数。因此，似然函数为

$$L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f\,(\varepsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\varepsilon'\varepsilon\right) \tag{7.31}$$

现在由于可以写出$\boldsymbol{\varepsilon}=\boldsymbol{y}-\boldsymbol{X\beta}$，所以似然函数变为

$$L(\boldsymbol{y},\boldsymbol{\beta},\sigma^2)=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right) \quad (7.32)$$

正如简单线性回归的情形，为方便处理，取似然函数的对数

$$\ln L(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2)=-\frac{n}{2}\ln(2\pi)-n\ln(\sigma)-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}) \quad (7.33)$$

显然对于定值σ，

$$(\mathbf{y}-\mathbf{X\beta})'(\mathbf{y}-\mathbf{X\beta})$$

项最小时，似然函数的值最大。因此，正态误差下$\boldsymbol{\beta}$的最大似然估计量等价于最小二乘估计量$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$。$\hat{\sigma}^2$的最大似然函数为

$$\tilde{\sigma}^2=\frac{(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n} \quad (7.34)$$

通俗的讲，拿到很多个样本（数据集中所有因变量）后，最大似然估计就是找到那个参数估计值，使得历史样本值发生概率最大。因收集的样本已发生，其发生概率最大才符合逻辑。求样本所有观测的联合概率最大化为连乘积，将其取对数变为线性加和。此时通过对参数求导数，并令一阶导数为零，就可以通过解方程（组），得到最大似然估计值。

7.3 高维回归系数压缩

随着复杂工业系统发展，可以获得反映多方面的数据，但很多数据存在强耦合等问题。多元线性回归模型中，若将全部数据用于回归分析，不仅导致问题难度增加，也容易造成过拟合使测试数据误差方差过大。因此减少不必要的特征，简化模型是减小方差的一个重要步骤。除了直接对特征筛选，也可以进行特征压缩，减少某些不重要的特征系数，系数压缩趋近于 0 就可以认为舍弃该特征。本章将从共线性出发，介绍不同的高维回归系数压缩方法。

7.3.1 共线性的来源及影响

特别将考虑多重共线性的起因，多重共线性对推断的某些特定影响，探测存在多重共线性的方法，以及处理多重共线性问题的一些方法。

（1）来源

写出多元回归模型

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{7.35}$$

式中：$\boldsymbol{y}$为$n \times 1$的响应变量向量，$\boldsymbol{X}$为$n \times p$向量，$\boldsymbol{\beta}$为未知常数的$p \times 1$向量，而$\boldsymbol{\varepsilon}$为随机误差的$n \times 1$向量。因此，$\boldsymbol{X'X}$为回归变量之间的相关系数的$p \times p$矩阵，而$\boldsymbol{X'y}$为回归变量与响应变量之间相关系数的$p \times 1$向量。

将$\boldsymbol{X}$矩阵的第j列记为$\boldsymbol{X_j}$，所以$\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_p}]$。因此，$\boldsymbol{X_j}$包含第$j$个回归变量的$n$个水平。可以根据$\boldsymbol{X}$列的线性相关性正式定义多重共线性。如果存在不全为零的常数集$t_1, t_2, \cdots, t_p$满足

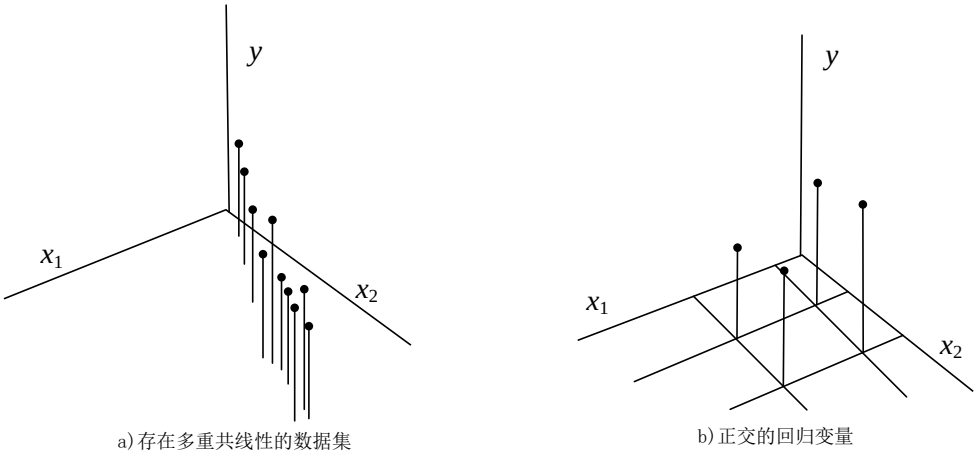
$$\sum_{j=1}^{p} t_j \boldsymbol{X_j} = 0$$

那么向量$\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_p}$是线性相关的，如果$\boldsymbol{X}$列的子集恰好满足方程，那么$\boldsymbol{X'X}$矩阵的秩会小于 p，同时$(\boldsymbol{X'X})^{-1}$不存在。多重共线性问题的程度是不同的，也就是说，除非$\boldsymbol{X}$的列是正交的($\boldsymbol{X'X}$为对角矩阵)，否则每个数据集都将遭受某种程度上的多重共线性。一般情况下，$\boldsymbol{X'X}$正交仅会发生在试验设计中。正如将要看到的，多重共线性的存在使得回归模型的常规最小二乘分析出现了极大的不适用性。

多重共线性主要有四种来源：1）解释变量都享有共同的时间趋势；2）一个解释变量是另一个的滞后，二者往往遵循一个趋势；3）由于数据收集的基础不够宽，某些解释变量可能会一起变动；4）某些解释变量间存在某种近似的线性关系。

（2）影响

多重共线性的存在对回归系数的最小二乘估计量有许多潜在的严重影响，其中某些影响是易于展示的。对图 7.3a）中的数据构建回归模型，类似于一个通过这些点的平面。显然这一平面将非常不稳定，同时对数据点相当小的变化敏感。进一步来说，模型可以在与样本中所观测的点类似的点处相当好地预测y，但是任何远离这一路径的外推法都可能产生不良的预测。作为对比，考察图7.3b）中的正交回归变散，这些点所拟合的平面将更为稳定。



a) 存在多重共线性的数据集　　b) 正交的回归变量

图 7.3 多重共线性数据与正交回归变量对比

假设只存在两个回归变量，假设x_1、x_2与y都已经尺度化为单位长度。模型为

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{7.36}$$

而最小二乘正规方程为

$$(\boldsymbol{X'X})\hat{\boldsymbol{\beta}} = \boldsymbol{X'y} \tag{7.37}$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} \tag{7.38}$$

式中：r_{12}为x_1与x_2之间的简单相关系数；r_{jy}为x_j与y之间的简单相关系数（j=1,2）。现在（$\boldsymbol{X'X}$）的逆为

$$\boldsymbol{C} = (\boldsymbol{X'X})^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix} \tag{7.39}$$

而回归系数的估计量为

$$\hat{\beta}_1 = \frac{r_{1y}-r_{12}r_{2y}}{1-r_{12}^2}, \hat{\beta}_2 = \frac{r_{2y}-r_{12}r_{1y}}{1-r_{12}^2} \tag{7.40}$$

如果x_1与x_2之间存在强烈的多重共线性，那么相关系数r_{12}将较大。由式

（7.38）可得，随着$|r_{12}| \to 1$, $Var(\hat{\beta}_j) = C_{jj}\sigma^2 \to +\infty$；而$Var(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \to \pm\infty$取决于$r_{12} \to +1$还是$r_{12} \to -1$。因此，$x_1$与$x_2$之间强烈的多重共线性会使得回归系数的最小二乘估计具有较大的方差与协方差。这意味着在相同的x水平下所采集的不同样本，产生的模型参数估计值可以有相当大的不同。

当存在多于两个的回归变量时，多重共线性会产生类似的影响。多重共线性往往会产生绝对值过大的最小二乘估计量$\hat{\beta}_j$。当存在强烈的多重共线性时，虽然最小二乘法一般会产生不良的单个模型参数估计值，但是这并不一定意味着拟合模型将做出不良的预测。

如果将预测控制在近似满足多重共线性的x空间区域内，那么拟合模型通常会产生令人满意的预测。即使对单个参数β_j的估计是不良的，因可以较好将线性组合$\sum_{i=1}^{p} \beta_j x_{ij}$估计，也会产生令人满意的预测。也就是说，如果原数据近似沿着上述所定义的超空间排布，那么尽管单个模型参数的估计值是不适用的，也可以精确地预测同样靠近这一超空间排布的未来观测值。下面用例子来更直观来说明多重共线性问题。

例 7.2 假设已知x_1，x_2与y的关系服从线性回归模型

$$y = 10 + 2x_1 + 3x_2 + \varepsilon$$

给定x_1，x_2的 10 个值，如下表 7-3。

表 7-3 x_1，x_2与y的值

序号	1	2	3	4	5	6	7	8	9	10
x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5
ε	0.8	-0.5	0.4	-0.5	0.2	1.9	1.9	0.6	-1.5	-1.5
y	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0

假设回归系数与误差项未知，采用最小二乘估计来求回归系数得，

	估计值	原模型
β_0	11.292	10
β_1	11.307	2
β_2	-6.591	3

易看出相差太大，通过计算x_1，x_2的相关系数易得为 0.986，表明x_1与x_2之间高度相关，即存在共线性问题，使得最小二乘估计不能满足当前模型。

但也要说明，不能存在多重共线性，不代表不能存在相关性（机器学习不要求特征之间必须独立）。在现实中特征之间完全独立的情况其实非常少，因为大部分数据统计手段或者收集者并不考虑统计学或者机器学习建模时的需求，现实数据多多少少都会存在一些相关性，极端情况下，甚至还可能出现收集的

特征数量比样本数量多的情况。

通常来说，这些相关性在机器学习中无伤大雅（在统计学中他们可能是比较严重的问题），即便有一些偏差，只要最小二乘法能够求解，都有可能会无视掉它。想要消除特征的相关性，不管使用什么方法都会进行特征选择，这意味着可用的信息变得更少，对于机器学习来说，排除相关性后，模型的整体效果会受到巨大的影响。

这种情况下，选择不处理相关性，然而多重共线性就不是这样一回事了，它的存在会造成模型极大地偏移，无法模拟数据的全貌，因此这是必须解决的问题。

7.3.2 岭回归法

将最小二乘法应用于非正交数据时，可能会得出非常不良的回归系数估计值。在 7.3.1 节看到回归系数估计值的最小二乘方差可能会有相当大的膨胀，最小二乘参数估计量的向量长度平均而言也会太长。这意味着最小二乘估计值的绝对值太大，也就非常不稳定，也就是说，最小二乘估计量的大小与正负号在给定不同样本时会有非常大的变化。

为了得到回归系数的有偏估计量，最初由 Hoerl 和 Kennard 提出，通过求解经过轻微修正的正规方程，可以求出岭估计量。岭回归损失函数的完整表达式为：

$$\min_{\widehat{\boldsymbol{\beta}}_{岭}} \left\| X\widehat{\boldsymbol{\beta}}_{岭} - y \right\|_2^2 + \lambda \left\| \widehat{\boldsymbol{\beta}}_{岭} \right\|_2^2 \tag{7.41}$$

求导易得：

$$\begin{aligned} \frac{\partial(RSS+\alpha\|\widehat{\boldsymbol{\beta}}_{岭}\|_2^2)}{\partial\widehat{\boldsymbol{\beta}}_{岭}} &= \frac{\partial(\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{岭}\|_2^2+\alpha\|\widehat{\boldsymbol{\beta}}_{岭}\|_2^2)}{\partial\boldsymbol{w}} \\ &= \frac{\partial(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{岭})'(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{岭})}{\partial\widehat{\boldsymbol{\beta}}_{岭}} + \frac{\partial\lambda\|\widehat{\boldsymbol{\beta}}_{岭}\|_2^2}{\partial\widehat{\boldsymbol{\beta}}_{岭}} \\ &= 0 - 2\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{2X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{岭} + \boldsymbol{2}\lambda\widehat{\boldsymbol{\beta}}_{岭} \end{aligned} \tag{7.42}$$

整理易得：

$$(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\widehat{\boldsymbol{\beta}}_{岭} = \boldsymbol{X}'y \tag{7.43}$$

当变量$|\boldsymbol{X}'\boldsymbol{X}| \approx 0$，即

$$\boldsymbol{X}'\boldsymbol{X} = \begin{vmatrix} \partial_{11} & \partial_{12} & \cdots & \partial_{1n-1} & \partial_{1n} \\ 0 & \partial_{22} & \cdots & \partial_{2n-1} & \partial_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \partial_{n-1n-1} & \partial_{n-1n} \\ 0 & 0 & \cdots & 0 & 0 \end{vmatrix}$$

通过加上$\lambda\boldsymbol{I}$，

$$\boldsymbol{X'X}+\lambda\boldsymbol{I}=\begin{vmatrix}\partial_{11}+\lambda & \partial_{12} & \cdots & \partial_{1n-1} & \partial_{1n}\\ 0 & \partial_{22}+\lambda & \cdots & \partial_{2n-1} & \partial_{2n}\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & \partial_{n-1n-1}+\lambda & \partial_{n-1n}\\ 0 & 0 & \cdots & 0 & \lambda\end{vmatrix}$$

保证$\boldsymbol{X'X}+\lambda\boldsymbol{I}$接近奇异程度小得多，只需要调整$\lambda$的值，即可保证矩阵永远满秩，即永远存在矩阵的逆。

当$\lambda\neq 0$ 时,$(\boldsymbol{X'X}+\lambda\boldsymbol{I})$可逆，故

$$\widehat{\boldsymbol{\beta}}_{\text{岭}}=(\boldsymbol{X'X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X'}y \tag{7.44}$$

式中$\lambda\geq 0$为所选择的常数。注意当$\lambda=0$时，岭估计量是最小二乘估计量。

由于

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{岭}}&=(\boldsymbol{X'X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X'}y=(\boldsymbol{X'X}+\lambda\boldsymbol{I})^{-1}(\boldsymbol{X'X})\widehat{\boldsymbol{\beta}}\\&=Z_k\widehat{\boldsymbol{\beta}}\end{aligned} \tag{7.45}$$

所以，岭估计量是最小二乘估计量的线性变换。

回顾最小二乘的表达式$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$，现在假设矩阵$\boldsymbol{X}$是列正交的，即$\boldsymbol{X'X}=I$。此时岭回归估计系数$\widehat{\boldsymbol{\beta}}_{\text{岭}}$的任一分量与最小二乘估计有如下对应关系：

$$\hat{\beta}_j^{ridge}=\frac{\hat{\beta}_j}{1+\lambda} \tag{7.46}$$

易看出，岭回归将最小二乘的系数缩小。

参数λ可以称为偏倚参数，当λ离开 0 时，估计的偏倚量增加。另一方面，回归系数估计的方差之和（称为总方差）为

$$\text{总方差}(\lambda)=\sum_{j=1}^{p}Var\left(\hat{\theta}_j(\lambda)\right)=\sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+\lambda)^2} \tag{7.47}$$

它是λ的递减函数公式，说明参数λ对于回归系数岭估计的总方差的影响。在式（7.47）中，$\lambda=0$时，得到

$$\text{总方差}(0)=\sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j} \tag{7.48}$$

这说明最小二乘估计的总方差受到小的特征值的影响是很大的。

当λ的值离开 0 趋于无穷时，所有的估计趋于零。岭估计的一个想法是选择适当的λ，使得总方差的减少量不超过偏倚的增加量。研究得出，存在一个参数λ的正值，使得岭估计对于数据中的小幅变动是相对稳定的。

岭回归的岭参数λ的选择：1）岭迹法；2）方差扩大因子法；3）残差平均和来确定λ值。此处仅做介绍，具体的可以自行学习。

例 7.3 近年来，钢铁产业一直保持高速发展。为了规划中国未来钢铁产业的发展，需要定量地分析影响中国市场发展的主要因素。经分析，影响国内市场收

入的主要因素，除了国内需求和国外需求以外，还可能与相关基础建设有关。为此，考虑的影响因素主要有国内消费数X_1，城镇建设需求X_2，农村建设需求X_3，并以X_4和X_5作为相关基础设施的代表。为此设定了如下对数形式的计量经济模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

具体数据，如表所示：

表 7-4 十年的钢铁业收入与相关数据

观测序号	响应变量	预测变量				
	Y	X_1	X_2	X_3	X_4	X_5
1	1023.5	52400	414.7	54.9	111.78	5.90
2	1375.7	62900	464.0	61.5	115.70	5.97
3	1638.4	63900	534.1	70.5	118.58	6.49
4	2112.7	64400	599.8	145.7	122.64	6.60
5	2391.2	69450	607.0	197.0	127.85	6.64
6	2831.9	71900	614.8	249.5	135.17	6.74
7	3175.5	74400	678.6	226.6	140.27	6.87
8	3522.4	78400	708.3	212.7	169.80	7.01
9	3878.4	87800	739.7	209.1	176.52	7.19
10	3442.3	87000	684.9	200.0	180.98	7.30

通过线性分析，由此可见，该模型R^2=0.9954，$\overline{R^2}$= 0.9897 可决系数很高，F 检验值 173.3525，明显显著。但是当$\alpha = 0.05$，不仅X_1、X_5系数的 t 检验不显著，而且X_5系数的符号与预期的相反，这表明很可能存在严重的多重共线性。

计算各解释变量的相关系数，得相关系数矩阵：

表 7-5 各相关系数矩阵

	X_1	X_2	X_3	X_4	X_5
X_1	1	0.9189	0.7520	0.9480	0.9417
X_2	0.9189	1	0.8651	0.8592	0.9633
X_3	0.7520	0.8651	1	0.6649	0.8181
X_4	0.9480	0.8592	0.6649	1	0.8977
X_5	0.9417	0.9633	0.8181	0.8977	1

通过相关系数矩阵易看出，各解释变量相互之间的相关系数较高，证实存在严重共线性问题。故采用岭回归解决该问题，最终拟合方程为：

$$Y = -2441.16 + 4.21X_3 + 3.22X_4 + 13.6X_5$$

当$\alpha = 0.05$，不仅X_3、X_4、X_5系数的 t 检验显著，这表明这是消除多重共线性的结果。

7.3.3 LASSO 回归法

虽然岭回归能提升模型拟合的精确度，但拟合的系数都非零。也就是说，岭回归达到变量选择的目的。LASSO（Least Absolute Shrinkage and Selection Operator）通过一阶惩罚项，能将一些系数恰好压缩为零，实现变量选择。除此之外，在高维问题，LASSO 有较好的预测精度和计算能力。LASSO 回归损失函数的完整表达式为：

$$\min_{\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}} \left\| X\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}} - y \right\|_2^2 + \alpha \left\| \widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}} \right\|_1 \tag{7.49}$$

求导易得：

$$\begin{aligned} \frac{\partial (RSS + \alpha \|\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}\|_1)}{\partial \widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}} &= \frac{\partial (\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}\|_2^2 + \alpha \|\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}\|_1)}{\partial \boldsymbol{w}} \\ &= \frac{\partial (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}})}{\partial \widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}} + \frac{\partial \alpha \|\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}\|_1}{\partial \widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}} \\ &= 0 - 2\boldsymbol{X}'\boldsymbol{y} + \mathbf{2}\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}} + \mathbf{2}\alpha \end{aligned} \tag{7.50}$$

整理易得：

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}} = \boldsymbol{X}'\boldsymbol{y} - \frac{\alpha \boldsymbol{I}}{\mathbf{2}} \tag{7.51}$$

LASSO 无法解决（岭回归可以解决）特征间精确相关关系导致的最小二乘无法使用的问题，即$\boldsymbol{X}'\boldsymbol{X}$不满秩。假设$\boldsymbol{X}'\boldsymbol{X}$的逆存在，则有

$$\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y} - \frac{\alpha \boldsymbol{I}}{\mathbf{2}}\right) \tag{7.52}$$

通过增大α，可以为$\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}$的计算增加一个负项，限制参数估计中$\widehat{\boldsymbol{\beta}}_{\mathbf{LASSO}}$的大小，从而防止多重共线性引起的参数被估计过大导致模型失准的问题。LASSO 并不是从根本上解决多重共线性问题，而是限制多重共线性带来的影响。

两个正则化都会压缩系数的大小，对标签贡献更少的特征的系数会更小，也会更容易被压缩。不过，L2 正则化（岭回归）只会将系数压缩到尽量接近 0，但 L1 正则化（LASSO 回归）主导稀疏性，因此会将系数压缩到 0。

结合图 7.4 模型的来进一步区分岭回归以及 Lasso 回归在求解最优解的差异，假设截距为 0 且$\boldsymbol{\beta} = (\beta_1, \beta_2)^T$为二维。

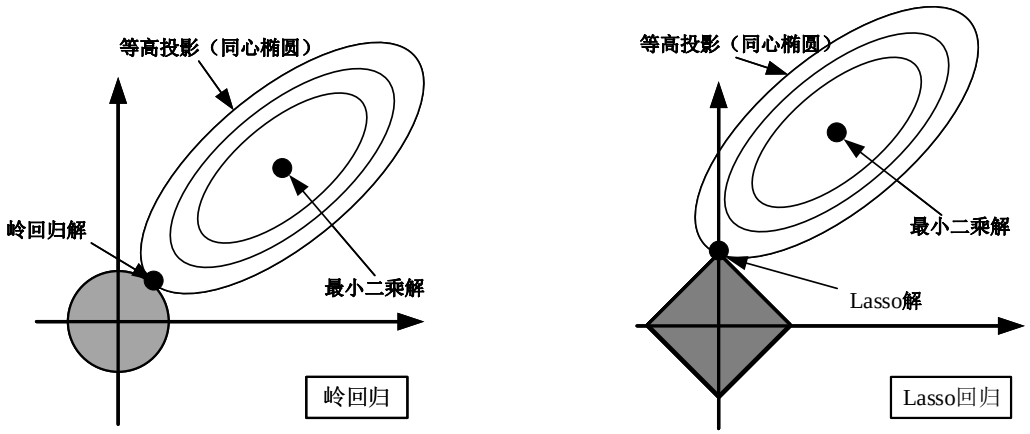


图 7.4 岭回归与 Lasso 回归最优解的区别

图 7.4 同心椭圆为最小二乘的解的等高投影，由于 Lasso 回归是 L1 范数，因此会出现“棱角”。当“棱角”与抛物面相交时，就会导致特征项易出现 0（正方形在这里斜率为±1，不像圆相切要求较为苛刻）。因此，相较于岭回归的 L2 范数通过降低各系数的绝对值而防止过拟合、降低模型复杂度而言，Lasso 回归的 L1 范数惩罚项易构造稀疏矩阵，有特征选择作用，同时也有一定程度防止过拟合的作用。

7.3.4 主成分回归法

本部分则介绍主成分回归，它不仅可以解决变量的共线性问题，还可以有效提取数据的特征，降低数据的冗余。在实际建模过程中面临的问题远不止共线性问题，更有维度灾难问题。在实际工业过程中，传感器测得特征个数成百上千，从而导致计算复杂度产生“指数爆炸”效应，但往往并不能给模型准确度带来提升。因此，利用 3.5.1 中提及的 PCA 对数据进行降维，即提取出数据中最具有代表性、包含波动信息最多的成分，并使用提取的主元特征来代替原预测变量进行建模。具体分析步骤参考本书 3.5.1。

模型的标准形式为

$$\boldsymbol{y}=\boldsymbol{Z\alpha}+\boldsymbol{\varepsilon} \tag{7.53}$$

式中

$$\boldsymbol{Z}=\boldsymbol{XT},\boldsymbol{\alpha}=\boldsymbol{T}'\boldsymbol{\beta},\boldsymbol{T}'\boldsymbol{X}'\boldsymbol{XT}=\boldsymbol{Z}'\boldsymbol{Z}=\boldsymbol{\Lambda}$$

$\mathbf{\Lambda}=\mathrm{diag}\left(\lambda_1,\lambda_2,\cdots,\lambda_p\right)$ 为 $\mathbf{X}'\mathbf{X}$ 特征值的 $p\times p$ 矩阵，而正交矩阵的列为 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 的特征向量。Z的列定义了一个新的正交回归变量，满足

$$\boldsymbol{Z}=\left[\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_p\right] \tag{7.54}$$

称之为主成分。

α的最小二乘估计量为

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{\Lambda}^{-1}\boldsymbol{Z}'\boldsymbol{y} \tag{7.55}$$

而α的协方差矩阵为

$$Var(\widehat{\boldsymbol{\alpha}}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \sigma^2\boldsymbol{\Lambda}^{-1} \tag{7.56}$$

主成分回归方法通过使用小于模型中主成分全集的集合来解决多重共线性问题。为了得到主成分估计量，假设回归变量是按其特征值的降序$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$排列的。假设这些特征值中的最后$s$个接近零。在主成分回归中，将接近零的特征值所对应的主成分从分析中移除，并将最小二乘应用于剩余的主成分。也就是说，

$$\widehat{\boldsymbol{\alpha}}_{\text{主成分}} = \boldsymbol{B}\widehat{\boldsymbol{\alpha}} \tag{7.57}$$

式中$b_1 = b_2 = \cdots = b_{p-s} = 1$且$b_{p-s+1} = b_{p-s+2} = \cdots = b_p = 0$。因此，主成分估计量为

$$\widehat{\boldsymbol{\alpha}}_{\text{主成分}} = \begin{bmatrix} \widehat{\boldsymbol{\alpha}}_{\mathbf{1}} \\ \widehat{\boldsymbol{\alpha}}_{\mathbf{1}} \\ \vdots \\ \widehat{\boldsymbol{\alpha}}_{\mathbf{p-s}} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{7.58}$$

以标准化回归变量为单位，即

$$\hat{\beta}_{\text{主成分}} = T\hat{\alpha}_{\text{主成分}} = \sum_{j=1}^{p-s} \lambda_j^{-1} t_j' X' y t_j \tag{7.59}$$

虽然主元分析相较于岭回归有所改进，但在讨论主元个数选取时，尽管包含了大部分的自变量信息，但并不能保证主成分与因变量的相关性。因此，如何提取既包含自变量数据信息，又与因变量保持较强关联的成分则成了进一步研究的方向。下一节的偏最小二乘回归则是基于此思想提出的。

7.3.5 偏最小二乘回归法

偏最小二乘回归最先产生于化学领域，在利用分光镜来预测化学样本的组成时，作为解释变量的红外区反射光谱的波长常有几百个，往往超过化学样本数的个数；所造成的多重相关性使得人们很难利用传统的最小二乘法。基于此，岭回归、主元回归等算法被相继提出。但易于发现主元回归是分裂的两步策略，即特征提取与回归模型的建立是不相关的。这导致提取的特征虽然概括了自变

量数据的大部分信息，却不能保证预测的准确性。此外，这种分裂的策略也导致了主元成分选取的困难。既然如此，是否可以直接提取因变量最相关的特征来建立回归模型呢？因此有学者提出了偏最小二乘回归。

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。其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。

考虑k个$y_1, y_2, \cdots, y_k$响应变量与p个预测变量$x_1, x_2, \cdots, x_p$的建模问题。偏最小二乘回归的基本思路是：首先在预测变量的变量集中选择第一成分u_1（u_1是$x_1, x_2, \cdots, x_p$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$y_1, y_2, \cdots, y_k$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$y_1, y_2, \cdots, y_k$与$u_1, u_2, \cdots, u_r$的回归式，然后再表示为$y_1, y_2, \cdots, y_k$与原自变量的回归方程式，即偏最小二乘回归方程式。

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选择第一对成分为u_1和v_1，u_1是自变量集$\boldsymbol{X} = [x_1, \cdots, x_p]^T$的线性组合

$$u_1 = \alpha_{11}x_1 + \cdots + \alpha_{1p}x_p = \boldsymbol{\rho_1^T X} \tag{7.60}$$

v_1是自变量集$\boldsymbol{Y} = [y_1, \cdots, y_k]^T$的线性组合

$$v_1 = \beta_{11}y_1 + \cdots + \beta_{1k}y_k = \boldsymbol{\gamma_1^T Y} \tag{7.61}$$

为了回归分析的需要，要求u_1与v_1各自尽可能多地提取所在变量组的变异信息，使得u_1与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$\boldsymbol{X}$和$\boldsymbol{Y}$，可以计算第一对成分的得分向量，记为$\hat{u}_1$和$\hat{v}_1$

$$\hat{u}_1 = A\rho_1 = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix} \begin{bmatrix} \alpha_{11} \\ \vdots \\ \alpha_{1p} \end{bmatrix} \tag{7.62}$$

$$\hat{v}_1 = B\gamma_1 = \begin{bmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{nk} \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \vdots \\ \beta_{1k} \end{bmatrix} \tag{7.63}$$

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，

$$\max\langle \hat{u}_1, \hat{v}_1 \rangle = \langle \mathrm{X}\rho_1, \mathrm{Y}\gamma_1 \rangle = \rho_1^T \mathrm{X}^T \mathrm{Y}\gamma_1 \tag{7.64}$$

$$\text{s.t.}\begin{cases}\rho_1^T\rho_1 = 1\\ \gamma_1^T\gamma_1 = 1\end{cases}$$

利用 Lagrange 乘数法，问题化为求单位向量ρ_1和γ_1，使$\theta_1 = \rho_1^T X^T Y\gamma_1$达到最大。问题的求解只须通过计算$p \times p$矩阵

$$\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X} \tag{7.65}$$

的特征值和特征向量，且$\boldsymbol{M}$的最大特征值为θ_1^2，相应的单位特征向量就是所求的解ρ_1，而γ_1可由ρ_1计算得到

$$\gamma_1 = \frac{1}{\theta_1}\boldsymbol{Y}^T\boldsymbol{X}\rho_1 \tag{7.66}$$

2）建立$y_1, y_2, \cdots, y_k$与u_1的回归及$x_1, x_2, \cdots, x_p$与u_1的回归

假定回归模型为

$$\begin{cases}\boldsymbol{X} = \hat{u}_1\boldsymbol{\sigma_1^T} + \boldsymbol{E_1}\\ \boldsymbol{Y} = \hat{u}_1\boldsymbol{\tau_1^T} + \boldsymbol{F_1}\end{cases} \tag{7.67}$$

其中$\sigma_1 = \left[\sigma_{11}, \cdots, \sigma_{1p}\right]^T, \tau_1 = [\tau_{11}, \cdots, \tau_{1k}]^T$分别是多对一的回归模型中的参数向量，$\boldsymbol{E_1}$和$\boldsymbol{F_1}$是残差阵。

3）用残差阵$\boldsymbol{E_1}$和$\boldsymbol{F_1}$代替 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 重复以上步骤。记$\hat{X} = \hat{u}_1\sigma_1^T$，$\hat{Y} = \hat{u}_1\tau_1^T$，则残差阵$\boldsymbol{E}_1 = \boldsymbol{X} - \widehat{\boldsymbol{X}}$，$\boldsymbol{F}_1 = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$。如果残差阵$\boldsymbol{E}_1$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的精度已满足需要了，可以停止选择变量。否则用残差阵$\boldsymbol{E_1}$和$\boldsymbol{F_1}$代替$\boldsymbol{X}$和$\boldsymbol{Y}$重复以上步骤即得。

$$\rho_2 = \left[\alpha_{21}, \cdots, \alpha_{2p}\right]^T \tag{7.68}$$

$$\gamma_2 = [\beta_{21}, \cdots, \beta_{2k}]^T \tag{7.69}$$

4）设$n \times p$数据阵$\boldsymbol{X}$的秩为$r \le \min(n-1, p)$，则存在r个成分$u_1, u_2, \cdots, u_r$，使得

$$\begin{cases}X = \hat{u}_1\sigma_1^T + \cdots + \hat{u}_r\sigma_r^T + E_r\\ Y = \hat{u}_1\tau_1^T + \cdots + \hat{u}_r\tau_r^T + F_r\end{cases} \tag{7.70}$$

把$u_k = \alpha_{k1}x_1 + \cdots + \alpha_{kp}x_p\ (k = 1,2,\cdots,r)$，代入$Y = u_1\tau_1^T + \cdots + u_r\tau_r^T$，即得$k$个因变量的偏最小二乘回归方程式

$$y_j = c_{j1}x_1 + \cdots + c_{jp}x_p, j = 1,2,\cdots,k \tag{7.71}$$

集多元线性回归分析、典型相关分析、主因子分析等方法于一体的偏最小二乘回归方法更适用于因子分解机（Factor Machine, FM）算法，去解决大规模稀疏矩阵中特征组合，也可以避免数据非正态分布、因子结构不确定性和模型不能识别等潜在问题。

7.4 非线性回归

线性回归模型提供了充分而灵活的分析框架，并满足了许多回归分析的需要。但是，线性回归模型并非适用于所有情形。在许多工业过程分析问题中，响应变量与预测变量是通过一个已知的非线性函数联系起来的，这就会产生非线性回归模型。将最小二乘法应用于非线性模型时，所得到的正规方程是非线性的，而一般情况下非线性方程是难以求解的。处理非线性问题的常规方法是通过迭代直接将残差平方和最小化。本节将描述非线性回归模型的参数估计，并展示如何对模型参数做出恰当的推断。

7.4.1 非线性回归模型

在线性回归模型并不适用的情形下，响应变量与回归变量之间的关系可能是微分方程或微分方程的解。通常情况下，模型有非线性的形式。举例来说，模型

$$y = \theta_1 e^{\theta_2 x} + \varepsilon \tag{7.72}$$

是关于未知参数θ_1与θ_2非线性的。这里使用符号θ代表非线性模型中的参数，以强调线性情形与非线性情形之间的区别。

一般情况下，将非线性回归模型写为

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \varepsilon \tag{7.73}$$

式中：$\boldsymbol{\theta}$为未知参数的$p \times 1$向量；ε为不相关的随机误差项，其$E(\varepsilon) = 0$且$Var(\varepsilon) = \sigma^2$。

分析非线性模型

$$y = f(x, \boldsymbol{\theta}) + \varepsilon = \theta_1 e^{\theta_2 x} + \varepsilon \tag{7.74}$$

期望函数关于θ_1与θ_2的导数为

$$\frac{\partial f(x,\boldsymbol{\theta})}{\partial \theta_1} = e^{\theta_2 x} \quad 和 \quad \frac{\partial f(x,\boldsymbol{\theta})}{\partial \theta_2} = \theta_1 x e^{\theta_2 x} \tag{7.75}$$

由于导数是未知参数θ_1与θ_2的函数，所以模型是非线性的。

7.4.2 非线性最小二乘

前文观察到线性回归的最小二乘法会涉及最小化最小二乘函数

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i - \left(\boldsymbol{\beta}_0 + \sum_{j=1}^{k} \boldsymbol{\beta}_j x_{ij} \right) \right]^2$$

因为是线性回归模型，所以当对$S(\boldsymbol{\beta})$做关于未知参数的微分并使导数等于零时，所得到的正规方程是线性方程，因此是易于求解的。

现在考虑非线性回归模型的情形。模型为

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_i (i = 1,2,\cdots,n) \tag{7.76}$$

现在式中$x_i' = [1, x_{i1}, x_{i2}, \cdots, x_{ik}](i = 1,2, \cdots, n)$。最小二乘函数为

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta})]^2$$

为了求解最小二乘估计量，必须对**方程**做关于$\boldsymbol{\theta}$的每个元素的微分。这将为非线性回归情形提供p个正规方程。正规方程为

$$\sum_{i=1}^{n} [y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta})] \left[\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} \right]_{\theta = \hat{\theta}} = 0 \ (j = 1,2, \cdots, p)$$

在非线性回归模型中，大方括号中的导数将是未知参数的函数。进一步而言，期望函数也是非线性函数，所以正规方程可能会非常难以求解。

考虑非线性回归模型

$$y = \theta_1 e^{\theta_2 x} + \varepsilon \tag{7.77}$$

这一模型的最小二乘正规方程为

$$\sum_{i=1}^{n} \left[y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i} \right] e^{\hat{\theta}_2 x_i} = 0$$

$$\sum_{i=1}^{n} \left[y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i} \right] \hat{\theta}_1 x_i e^{\hat{\theta}_2 x_i} = 0$$

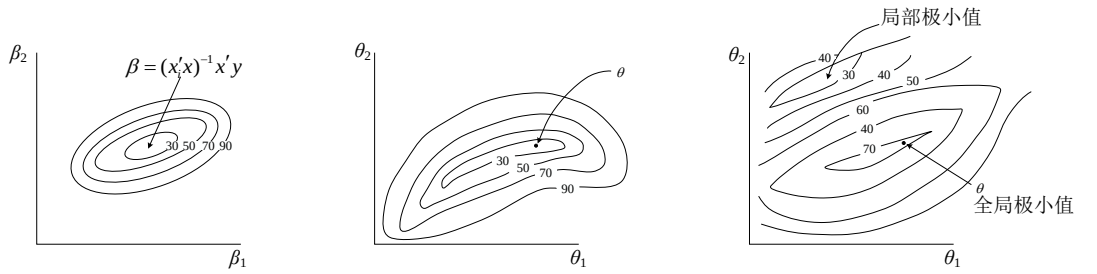
化简后，正规方程为

$$\sum_{i=1}^{n} y_i \, e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^{n} e^{2\hat{\theta}_2 x_i} = 0$$

$$\sum_{i=1}^{n} y_i \, x_i e^{\hat{\theta}_2 x_i} - \hat{\theta} \sum_{i=1}^{n} x_i \, e^{2\hat{\theta}_2 x_i} = 0$$

这两个方程不是$\hat{\theta}_1$与$\hat{\theta}_2$的线性方程，所以不存在简单的闭合形式解。一般情况下，必须使用迭代方法来求解$\hat{\theta}_1$与$\hat{\theta}_2$的值。使得问题进一步复杂化的是，正规方程有时会存在多个解，即残差平方和函数$S(\boldsymbol{\theta})$会存在多个平稳值。

当模型为非线性回归模型时，等高线通常如图 7.5，注意图 7.5 b)中的等高线不是椭圆，而事实上其形状是被严重拉长而不规则的，“香蕉形”的形状是非常典型的。残差平方和等高线的特定形状与方向取决于非线性模型的形式与所获得的数据样本。通常情况下靠近最优值时曲面会被严重拉长，所以$\boldsymbol{\theta}$的许多解产生的残差平方和都会接近于全局最小值。这会产生病态问题，而存在病态问题时通常难以求解$\boldsymbol{\theta}$的全局最小值。在某些情形下，等高线可能非常不规则，以至于会存在若干个局部极小值，也可能会存在多于一个的全局最小值$\boldsymbol{\theta}$。图 7.5 c)展示了存在一个局部极小值与一个全局最小值的情形。



a) 线性模型　　b) 非线性模型　　c) 有局部极小值与全局极小值的非线性模型

图 7.5 残差平方和函数的等高线图

如果模型中的误差项为独立正态分布且有常数方差，那么采用最大似然方法来对问题进行估计将会得到最小二乘法的估计结果。

如果误差为独立正态分布且均值为零，方差为σ^2，那么似然函数为

$$L(\theta,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{n/2}}exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i-\theta_1e^{\theta_2x_i}\right]^2\right]$$

显然，最大化这一似然函数等价于最小化残差平方和。因此，在正态理论中，最小二乘估计值与最大似然估计值相同。

非线性模型与线性模型的转换：变换会将线性引入模型期望函数，所以变换有时是有用的。举例来说，考虑模型

$$y=f(x,\boldsymbol{\theta})+\varepsilon=\theta_1e^{\theta_2x}+\varepsilon \tag{7.78}$$

由于$E(y)=f(x,\boldsymbol{\theta})=\theta_1e^{\theta_2x}$，所以可以使用对数将期望函数线性化：

$$\ln E(y)=ln\,\theta_1+\theta_2x \tag{7.79}$$

因此将模型重写为

$$\ln y=ln\,\theta_1+\theta_2x+\varepsilon=\beta_0+\beta_1x+\varepsilon \tag{7.80}$$

使用简单线性回归模型来估计β_0与β_1。但是，式（7.80）中参数的线性最小二乘估计量在一般情况下并不等价于原模型中的非线性参数估计量；其原因在于：在原非线性模型中最小二乘意味着最小化y的残差平方和，然而在变换后模型中最小化的是$\ln y$的残差平方和。

注意式（7.79）中的误差结构是加性，所以使用对数变换不可能会产生式（7.80）中的模型。而如果误差结构是乘性的，比如说

$$y=\theta_1e^{\theta_2x}\varepsilon \tag{7.81}$$

那么使用对数变换将是合适的，这是因为

$$\ln y=\ln\theta_1+\theta_2x+\ln\varepsilon=\beta_0+\beta_1x+\varepsilon^* \tag{7.82}$$

而如果ε^*服从正态分布，那么所有标准线性回归模型的性质与相关推断都可以应用进来。可以变换为等价线性形式的非线性模型称为非本质线性模型。

常见的变换如下表所示：

表 7-6 曲线方程与线性方程的变换

曲线方程	变换公式	变换后的线性方程
$\frac{1}{y}=a+\frac{b}{x}$	$X=\frac{1}{x}$，$\mathrm{Y}=\frac{1}{y}$	$Y=a+bX$
$y=ax^b$	$X=\ln x$，$\mathrm{Y}=\ln y$	$Y=a'+bX(a'=\ln x)$
$y=a+b\ln x$	$X=\ln x$，$\mathrm{Y}=y$	$Y=a+bX$
$y=ae^{bx}$	$X=x$，$\mathrm{Y}=\ln y$	$Y=a'+bX(a'=\ln x)$
$y=ae^{\frac{b}{x}}$	$X=\frac{1}{x}$，$\mathrm{Y}=\ln y$	$Y=a'+bX(a'=\ln x)$

7.4.3 支持向量回归

前边章节已经介绍过 SVM，它一般用于分类任务，本次我们所提到的支持向量回归（SVR）则是 SVM 在回归分析中的应用。

在 SVM 中我们的目标是通过最大化间隔，找到一个分离超平面，使得绝大多数的样本点位于两个决策边界的外侧。SVR 同样是考虑最大化间隔，但是考虑的是决策边界内的点，使尽可能多的样本点位于间隔内。

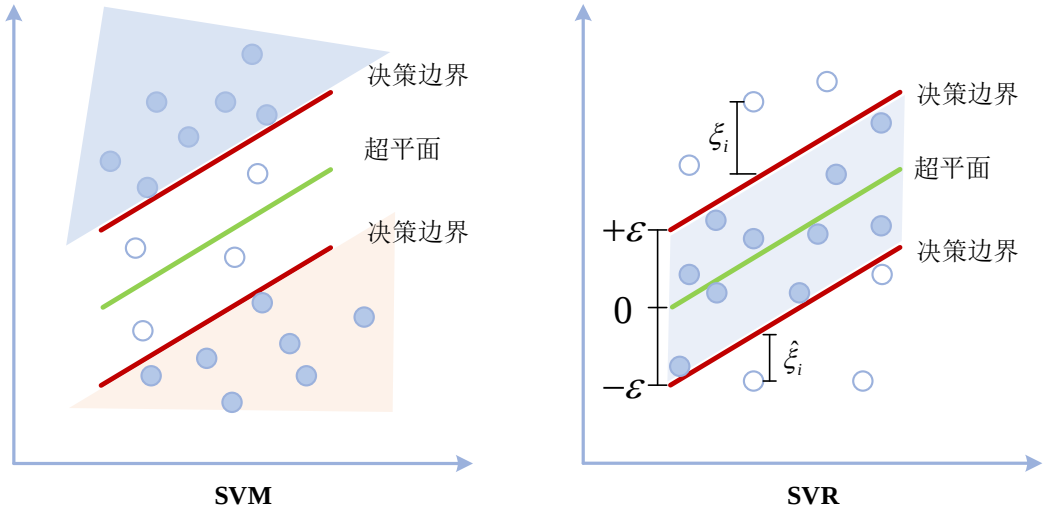


图 7.6 SVM 和 SVR 示意图

针对 SVR 的优化问题，为每个样本点引入松弛变量ξ_i与$\hat{\xi}_i$：

$$\min_{w}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{N}(\xi_i+\hat{\xi}_i)$$

$$\text{s.t. } y_i-\boldsymbol{w}^T\boldsymbol{x}_i-b\le\varepsilon+\xi_i$$

$$\boldsymbol{w}^T\boldsymbol{x}_i+b-y_i\le\varepsilon+\hat{\xi}_i$$

$$\xi_i,\hat{\xi}_i\ge 0$$

从上面的优化问题我们可以看出，SVR 只对间隔外的样本进行惩罚，当样

本点位于间隔内时，则不计算其损失。

对于非线性 SVR，自然地，引入核函数（kernel function）即可。

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (\alpha_i - \hat{\alpha}_i) \langle \varphi(\boldsymbol{x}_i), \varphi(\boldsymbol{x}) \rangle + b$$

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (\alpha_i - \hat{\alpha}_i) K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

常用的核函数有多项式（Polynomial）核$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^d$，高斯径向基函数（Gaussian Radial Basis function）$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j)\|^2/2\sigma^2)$等。

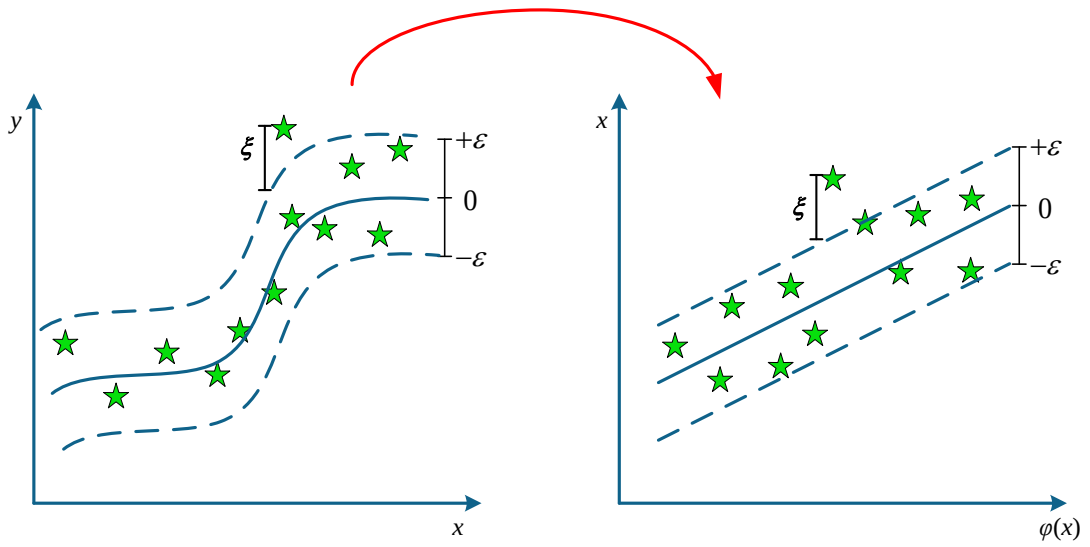


图 7.7 非线性与线性问题的转化示意图

对于 b 的计算，步骤如下：

1）从 KKT 等式 3,4 可以看出，如果ξ_i或$\hat{\xi}_i$不为零时，那么$C = \alpha_i$或$C = \hat{\alpha}_i$，也就是说该样本点位于间隔区间之外。

2）因为$y_i - f(\mathbf{x}_i) - \varepsilon - \xi_i = 0$与$f(\mathbf{x}_i) - y_i - \varepsilon - \hat{\xi}_i = 0$不可能同时成立（否则$\xi_i = \hat{\xi}_i = \varepsilon$显然不可能），因此$\alpha_i$和$\hat{\alpha}_i$不可能同时非零，即$\alpha_i \hat{\alpha}_i = 0$。

3）如果$0 < \alpha_i < C$，那么$\xi_i = 0$，同时$y_i - \boldsymbol{w}^T \boldsymbol{x}_i - b - \varepsilon - \xi_i = 0$，那么

$$b^* = y_i - \boldsymbol{w}^{*T} \boldsymbol{x}_i - \varepsilon \tag{7.86}$$

也就是说我们只要取$\alpha_i^* \in (0, C)$对应的样本点（称为支持向量）计算b^*即可。一般也可以取多个这样的点计算多个b值，再取平均作为b^*。同样当$\hat{\alpha}_i^* \in (0, C)$时，也可以用于计算

$$b^* = y_i - \boldsymbol{w}^{*T}\boldsymbol{x}_i + \varepsilon \tag{7.87}$$

对此感兴趣的读者可以查阅相关资料深入学习。

7.5 回归模型的验证

回归模型广泛用于预测与估计、数据描述、参数估计及控制。在将模型发布之前，应当评定模型验证，以区分模型适用性检验与模型验证。模型适用性检验包括残差分析、失拟检验、寻找高杠杆观测值与强影响观测值等。而模型验证所针对的是在运作环境中，模型是否会成功运行。

恰当的回归模型验证应当包括研究回归系数，以确定其正负号与大小是否是合理的。三种类型的程序对回归模型的验证是有用的：

1）模型系数与预测值的分析，包括与先验经验，实际理论，以及对其他分析模型或模拟的结果进行对比。

2）使用所收集的新数据来研究模型的预测性能。

3）数据分割，也就是说，拿出一部分原数据并使用这些数据的观测值来研究模型的预测性能，模型最终用途通常会表明合适的模型验证方法。因此，打算作为预测方程使用的模型验证应当集中于确定模型的预测精确性。

7.5.1 模型系数与预测值的分析

在研究最终回归模型的系数中，可通过对模型进行分析，以确定这些系数是否稳定以及这些系数的正负号与大小是否符合先验经验，获得关于回归模型影响的方向与相对大小的信息。将所估计模型的系数与这些信息进行对比，正负号不符合期望或绝对值过大的系数通常会表明要么模型是不适用的，要么单个回归变量影响的估计值是不良的。其他多重共线性诊断量也同样需要进行模型验证，如果有方差膨胀系数（Variance Inflation Factor，VIF）超过 5 或 10，那么因为存在回归变量之间的近线性关系，所以该特定回归系数是因估计不良而导致不稳定。当数据是跨时间收集的时，可以通过对较短的时间间隔拟合模型来考虑系数的稳定性。举例来说，如果有若干个年份的月度数据，对每个年份构建模型后，所希望的是每个年份的系数都是相似的。

响应变量的预测值可反映模型验证的度量。不切实际的预测值，比如正数数量的负数预测值或落在所期望响应变量范围之外的预测值，会表明所估计的系数是不良的，或者模型的形式是不正确的。

7.5.2 收集新数据验证

度量回归模型预测性能验证的最有效方法为：收集新数据并直接对比预测值与新观测值。如果模型对新数据给出了精确的预测，那么对模型与模型构建过程将都会有更强的信心，而这些新观测值有时称为确认性试验。预测性能的可靠性评定，设计两个或更多其他回归模型利用新数据对比这些模型的预测性能，可以为最终模型的选择提供参考。

7.5.3 数据分割

在许多情况下，为了达到数据验证的目的去收集新数据是不可行的。例如数据收集的预算可能已经花光了，可能已经转变为生产其他产品，或者不能获得收集数据所需要的其他设备或资源。当发生这些情况时，合理的方法是将可用数据分割为两个部分，分别为估计性数据（训练集）与验证性数据（测试集）。估计性数据用于构建回归模型，而后预测性数据用于研究模型的预测能力，也称为交叉验证。

如果所收集的数据是时间序列，那么时间就可以用作数据分割的基础。即寻找出特定的时间时期，将在这一时期之前所收集的观测值当作估计性数据集，而之后收集的观测值当作预测性数据集，而这种数据验证方法在时间序列分析中相当普遍。

除了时间之外，数据的其他特征通常也可以用于数据分割。在不存在数据分割的逻辑基础的情况下，可以随机安排观测值进入估计性数据集与预测性数据集。如果使用了随机安排，那么应当将这一随机安排过程重复若干次，使观测值的不同子集用于模型拟合。

7.5.4 模型拟合度量

模型是否适用与建模的目的紧密相关，所以很难得出统一的检验方法，而是要根据问题的性质采取不同的方法。一般来说，适用性检验在得到模型后进行，但也可以在回归过程的各个阶段进行。

（1）残差图

残差定义为

$$e_i = y_i - \hat{y}_i \ (\mathrm{i} = 1,2,\cdots,\mathrm{n})$$

式中：y_i为观测值；$\hat{y}_i$为所对应的拟合值。由于残差可以看作数据与拟合值之间的离差，所以残差也是响应变量中回归模型所未解释的变异性的度量。残差分析是探索模型适用性的有效方法。

残差的图形分析是研究回归模型拟合适用性与检验基本假设较为有效的方法，应当在所有回归建模问题中考虑残差图。

正态概率图：稍微违反正态性假设不会严重影响模型，但是由于t统计量或F统计量以及置信区间与预测区间依赖于正态性假设，所以总体上的非正态性可能更加严重。进一步来说，如果误差来自厚尾分布即重尾分布而不是正态分布，那么最小二乘拟合可能对数据的小型子集是敏感的。厚尾的误差分布通常会产生离群点，而离群点将最小二乘拟合过度地“拉”向离群点自身的方向。

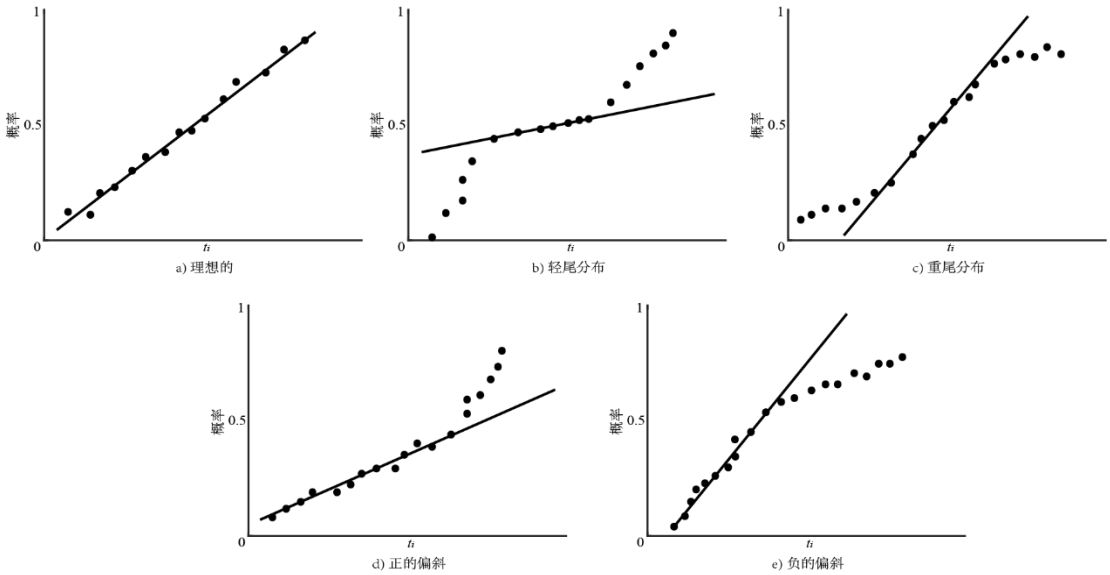


图 7.8 正态概率图

残差与拟合值$\hat{y}_i$的残差图：残差与对应拟合值$\hat{y}_i$的残差图对探测几种类型模型不适用性很有用。如果残差图类似与图 7.9 a)，它表明残差包含在一条水平带中，那么模型不存在明显的缺点。与$\hat{y}_i$的残差图类似于图 7.9 b）~图 7.9 d）中的任何模式都是模型存在缺点的表现。

图 7.9 b）与图 7.9 c）中的误差的方差不是常数。处理方差不相等通用的方法是对回归变量或响应变量二者之一使用合适的变换，或是使用加权最小二乘法。在实践中，通常利用对响应变量变换来得到稳定的方差。

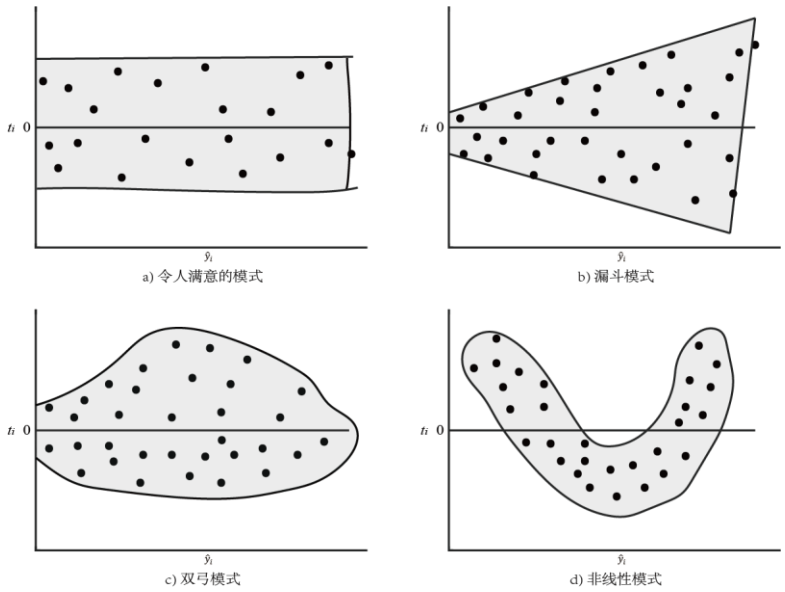


图 7.9 残差图的模式

图 7.9 d）中的曲线点表明存在非线性，这可能意味着模型需要其他回归变量。举例来说，平方项可能是必要的。对回归变量与响应变量之一或两者的变换在这种情形中可能也是有帮助的。

残差与$\hat{y}_i$的点图可能会解释一个或更多异常大的残差。这些点当然可能是离群点，出现在极端$\hat{y}_i$处的较大残差也可能表明要么方差不是常数，要么 y 与 x 之间的真实关系不是线性的。在考虑该点为离群点之前，应当先研究这几种可能的情况。

残差与回归变量的残差图：作出残差与每个对应回归变量值的残差图也是有益的。这种残差图通常展示了诸如图 7.9 的那些模式，除了水平尺度为第j个回归变量x_{ij}而不是$\hat{y}_i$。这里也是希望得到残差是包含在水平带中。图 7.9 b)中的漏斗模式与图 7.9 c)中的双弓模式都表明了方差非常数非常数方差（Non-Constant in Variance）。图 7.9 d)中的曲线带即非线性模式一般情况下意味着所假设的y与回归变量x_j之间的关系是不正确的。因此，应当考虑要么添加x_j的高阶项（如x_j^2），要么进行变换。

时间序列的残差图：如果已知所收集的数据为时间序列，那么做出残差与时间序列的残差图是个好方法。理想情况下，这一残差图将类似于图 7.9 a)，也就是说，一个水平带囊括了所有残差，而残差将在这一水平带中或多或少地以随机方式波动。但是，如果这张图类似于图 7.9 b)~图 7.9 d)，那么这可能表明方差随着时间变化，即应当向模型中添加时间的线性项或二次项。

残差的时间序列图可能会表明某一段时段上的误差与其他时段上的误差相关。不同时段上模型误差的相关性称为自相关：诸如图 7.10 a)的残差图表明存在正的自相关，而图 7.10 b)是负自相关的典型。

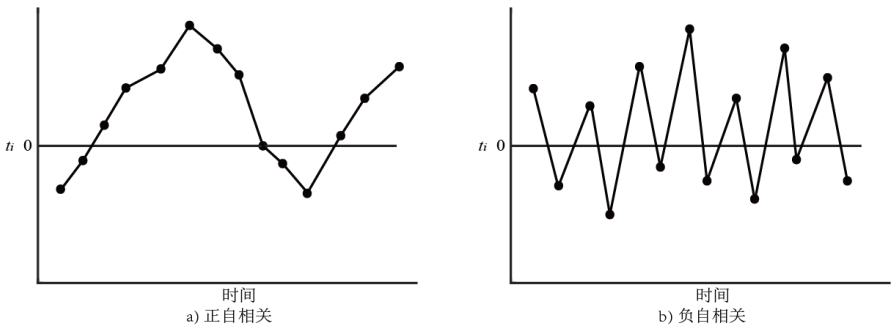


图 7.10 与时间的原型残差图，展示了误差的自相关

（2）杠杆率图

杠杆率图（leverage plot）也被称为偏杠杆率图（partial leverage plot）、偏残差图（partial residual plot）或者偏回归图（partial regression plot）。其本质是残差与预测值的残差图的变体，是加强对给定模型中其他回归变量时回归变量边界关系研究的一种方式。杠杆率图对响应变量y和回归变量x_j两个变量与模型

中其他回归变量和通过回归得到的每个残差作回归。这种残差之间相互的图形提供了关于所考虑回归变量边际关系的信息。

（3）拟合效果度量

在拟合了Y关于X的线性模型之后，我们不但想知道这种线性关系是否真的存在，还想度量模型对数据的拟合效果。拟合效果可以采用下面的方法之一进行度量，这些方法有很高的关联性。

1）考察Y对$\hat{Y}$的散点图，散点图上的这组点离一条直线越近，Y与X之间的线性关系越强。我们也可以通过计算Y和$\hat{Y}$的相关系数$\mathrm{Cor}(Y,\hat{Y})$来度量线性关系的强度，其中$\bar{y}$是响应变量Y的均值，$\bar{\hat{y}}$是拟合值$\hat{Y}$的均值。

2）另一个度量线性模型对观测数据拟合效果的非常有用。当我们获得线性模型参数的最小二乘估计后，再计算下面的量

$$\begin{aligned}\mathrm{SST}&=\sum(y_i-\bar{y})^2\\ \mathrm{SSR}&=\sum(\hat{y}_i-\bar{y})^2\\ \mathrm{SSE}&=\sum(y_i-\hat{y}_i)^2\end{aligned}\tag{7.88}$$

其中 SST 是Y偏离其均值$\bar{y}$的总离差平方和，SSR 是回归平方和，SSE 为残差平方和。在简单线性回归和多元线性回归中，下面的基本等式都是成立的

$$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE}\tag{7.89}$$

关于Y的总离差平方和 SST 可以分解为 SSR 和 SSE 两部分之和，其中 SSR 度量了X对Y的预测能力，SSE 度量了预测误差。因此，比值R^2 =SSR/SST 表示的是响应变量Y的总离差平方和中由预测变量X解释的比例。可将R^2表示为

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}\tag{7.90}$$

由 SSE≤SST，知$0\le R^2\le 1$。R^2是一个拟合优度指数，如果R^2接近 1，说明Y的绝大部分变化可由X解释。因此，R^2称为决定系数，反映了预测变量对响应变量的解释能力。在多元回归分析中，R^2也有类似的意义。

7.6 本章小结

回归分析是一组统计过程，用于估计因变量和一个或者多个自变量值之间的关系。本章首先介绍回归分析的概念，引出最常见、最简单的最小二乘回归法的基本推导步骤，并针对其无法解决多重共线性问题，分别介绍了岭回归、Lasso 回归等方法。其次针对非线性问题进行了分析，最后介绍了模型验证的方法。具体内容如下：

基本概念：介绍了回归分析的概念，并描述回归分析的基本步骤。对问题进行描述，选择相关变量，并收集相关数据，针对提出的问题来选择对应的拟合方法及其模型，最终解决问题。

线性回归：介绍了基本的线性回归模型，对线性模型进行分析，介绍了最小二乘估计法，对相关参数完成估计，并介绍了模型适用性检测，针对出现的问题，介绍了对应的修正方法（加权最小二乘方法）。

高维回归系数压缩：介绍了高维系数存在的共线性问题及其影响，并介绍对应的解决方法（岭回归、Lasso 回归、主成分回归、偏最小二乘回归）。

非线性回归：通过对比线性模型对非线性模型进行分析，介绍了非线性最小二乘估计法，对相关参数完成估计，并介绍了支持向量回归，通过引入核函数来解决非线性问题。

模型验证：通过划分训练集和测试集等方法，对拟合后的模型的效果进行检测，并修正模型的参数。

习题

7-1. 回归分析是怎样的一种统计方法，用来解决什么问题？

7-2. 思考并谈论回归分析与相关分析的异同。

7-3. 时间序列自身相关意义是什么？

7-4. 简要描述最小二乘法的几何意义。

7-5. 某钢铁厂某设备使用年限x和该年支出维修费用y（万元），数据如下：

使用年限x	2	3	4	5	6
维修费用y	2.2	3.8	5.5	6.5	7.0

(1) 求线性回归方程；

(2) 由（1）中结论预测第 10 年所支出的维修费用。

7-6. 下边为某年 5 地区的国内人均 GDP 和人均消费水平的统计数据：

地区	人均 GDP/元	人均消费水平
北京	22460	7326
辽宁	11226	4490
上海	34547	11546
河南	5444	2396
贵州	2662	1608

(1) 求人均 GDP 作自变量，人均消费水平作因变量，绘制散点图，并说明二者之间的关系形态。

(2) 计算两个变量之间的线性相关系数，说明两个变量之间的关系强度。

(3) 求出估计的回归方程,并解释回归系数的实际意义。

(4) 求人均 GDP 为 5000 元时，人均消费水平 95%的置信区间和预测区间。

(5) 如果某地区的人均 GDP 为 5000 元,预测其人均消费水平。

7-7. 请简单描述一下自变量$x_1, x_2, \cdots, x_n$之间存在多重共线性的定义；

7-8. 举出回归模型中使用正则化的几个例子，并总结正则化在不同情况下的作用。

7-9. 总结线性回归的优点。

7-10. 回归分析模型的种类及应用场景。

7-11. 思考上章内容，简述 SVM 用于分类和回归的联系与差别。

参考文献

[1] Montgomery D C. 等著; 王辰勇译. 线性回归分析导论[M]. 北京: 机械工业出版社, 2016. 4.

[2] Chatterjee S, Hadi A S. 著; 郑忠国, 许静译．例解线性回归[M]. 北京: 机械工业出版社，2013.8.

[3] 王星等. 大数据分析：方法与应用[M]. 北京: 清华大学出版社， 2013.

[4] Freund R J, Wilson W J, Sa P. Regression analysis[M]. Elsevier, 2006.

[5] Hoerl A E, Kennard R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.

[6] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(1): 53-71.

[7] Farrar D E, Glauber R R. Multicollinearity in regression analysis: the problem revisited[J]. The Review of Economic and Statistics, 1967: 92-107.

[8] Jolliffe I T. A note on the use of principal components in regression[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1982, 31(3): 300-303.

[9] Vinzi V E, Chin W W, Henseler J, et al. Handbook of partial least squares[M]. Berlin: Springer, 2010.

[10] Chen X, Cao W H, Gan C, Wu M. A hybrid partial least squares regression-based real time pore pressure estimation method for complex geological drilling process[J]. Journal of Petroleum Science and Engineering, 2022 210: 109771.

[11] Yuan X, Ge Z, Huang B, Song Z and Wang Y. Semisupervised JITL Framework for Nonlinear Industrial Soft Sensing Based on Locally Semisupervised Weighted PCR[J], IEEE Transactions on Industrial Informatics, 2017, 13(2):532-541.

[12] Yuan X, Li L and Wang Y. Nonlinear Dynamic Soft Sensor Modeling With Supervised Long Short-Term Memory Network[J], IEEE Transactions on Industrial Informatics, 2020, 16(5):3168-3176.