理论介绍

本书涉及的理论基础包括：区间规划模型、SLP 模型、模糊综合评价模型、分数阶理论以及数据挖掘等理论方法，将在本章进行详细的介绍及分析，为本书的理论研究奠定基础.

1.1 线性回归模型

回归分析是考察变量之间统计联系的一种重要方法，它在数学建模、自然科学和社会科学等许多领域中都有极其广泛的作用. 本节主要讨论一个随机变量或多个随机变量之间的关系. 回归(regression)一词是英国著名人类学家和气象学家 Francis Galton(1822—1911 年)于 1885 年引入的. 在“身高遗传中的平庸回归”的论文中，Galton 阐述了他的重大发现：虽然高个子的先代会有高个子的后代，但子代的身高并不像其父代，而是趋向于比他们的父代更加平均，就是说如果父亲身材高大，则子代的身材要比父代矮小一些；如果父亲身材矮小，则子代的身材要比父代高大一些. 因此，他用回归一词来描述子代身高与父代身高的这种关系. 而后，他的朋友、英国著名统计学家 K. Pearson 等人搜集了上千个家庭成员的身高数据，分析出儿子的身高 y 与父亲的身高 x 大致可归结为以下关系：

$$y = 0.516x + 33.73 \quad (\text{单位为英寸})$$

此式进一步证实了 Galton 的“回归定律”. 这就是回归一词最初在遗传学上的含义.

1.1.1 回归的概念

变量之间的关系大致可分为两类：一类是确定性的关系，如我们熟知的函

数关系；另一类是非确定性的关系. 对于某些非确定性的关系，如随机变量 Y 与变量 x（它可能是多维向量）之间的关系，当自变量 x 确定之后，因变量 Y 的值并不随着确定，而是按一定的统计规律（即随机变量 Y 的分布）取值. 这时我们将它们之间的关系表示为

$$Y = f(x) + \varepsilon$$

其中，$f(x)$ 是一个确定的函数，称为回归函数，ε 为随机项，且 $\varepsilon \sim N(0, \delta^2)$.

回归分析的任务之一是确定回归函数 $f(x)$. 当 $f(x)$ 是一元线性函数时，称为一元线性回归；当 $f(x)$ 是多元线性函数时，称为多元线性回归；当 $f(x)$ 是非线性函数时，称为非线性回归.

1.1.2 一元线性回归

1. 一元线性回归的数学模型

假定只研究 x 与 y 的关系，可以有如下结构式：

$$y = \beta_0 + \beta_1 x + \varepsilon$$

式中，β_0 和 β_1 是未知常数，ε 表示其他随机因素对 y 获得率的影响，它服从 $N(0, \sigma^2)$ 分布.

取定一组不完全相同的值 $x_1, x_2, \cdots, x_n$，作独立试验得到 n 对观察结果 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，其中 y_i 是 $x = x_i$ 处对随机变量 y 观察的结果. 将数据点 $(x_i, y_i)(i = 1, 2, \cdots, n)$ 代入，则有

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

并且假定 $\varepsilon_i \sim N(0, \sigma^2)$.

回归分析的首要任务是通过观察结果来确定回归系数 β_0, β_1 的估计 $\hat{\beta}_0, \hat{\beta}_1$，以下用最小二乘法确定回归直线方程：

$$y = \beta_0 + \beta_1 x$$

因为我们要确定一条直线，也就是要确定 β_0, β_1 的估计值，使回归直线与所有数据点都比较“接近”. 为了刻画这种“接近”程度，我们引进残差的概念，所谓残差是指观察值 y_i 与回归值 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 的偏差 $y_i - \hat{y}_i$，很自然地可以用绝对残差和 $\sum_{i=1}^{n} | y_i - \hat{y}_i |$ 来度量观察值与回归直线的接近程度. 绝对残差和越小，回归直线就与所有数据点越接近. 但为计算方便，一般用残差平方和：

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

来描述所有观察值与回归直线的偏离程度.

所谓最小二乘估计,就是使残差平方和 Q 达到最小值的 β_0,β_1 作为回归系数的估计.

令 $\frac{\partial Q}{\partial \beta_0}=0,\frac{\partial Q}{\partial \beta_1}=0$,得

$$\hat{\beta}_1=\frac{S_{xy}}{S_{xx}},\quad \hat{\beta}_0=\bar{y}-\beta_1\bar{x}$$

其中,

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2,S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

2. 一元线性回归的显著性检验

从前面求回归方程的过程来看,对任意样本观察值 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 做出的散点图,即使一看就知道这些点不可能近似在一条直线的附近,即 y 与 x 不存在线性关系,但是,若用最小二乘法仍然可求得 y 对 x 的线性回归方程与 $y=\beta_0+\beta_1 x$,这样求得的方程是没有意义的. 所以,在求 y 对 x 的线性回归方程之前,必须判断 y 与 x 的关系是否满足一元线性回归模型.

在处理具体问题时,判断 y 与 x 的关系是否满足一元线性回归模型,专业知识是重要的,在数学上,做出散点图是一种粗略的判断. 下面介绍根据样本观察值 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 进行统计检验的一种做法.

由 $y=\beta_0+\beta_1 x+\varepsilon$ 可知,当 $|\beta_1|$ 越大,y 随 x 的变化而变化的趋势就越明显;当 $|\beta_1|$ 越小,y 随 x 的变化而变化的趋势就不明显,特别当 $\beta_1=0$ 时,就认定 y 与 x 之间不存在线性相关关系. 这样,判断 y 与 x 是否满足一元线性回归模型就转化为在显著性水平 α 下,检验假设:

$$H_0:\beta_1=0,\quad H_1:\beta_1\neq 0$$

记 $S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2,S_{xy}=\sum_{i,j=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),U=\sum_{i=1}^{n}(\beta_0+\beta_1 x-\bar{y})^2=\frac{S_{xy}}{S_{xx}}$(称为回归平方和,$U+Q=S_{yy}$).

选取检验统计量:

$$F=\frac{U}{Q/(n-2)}$$

在 H_0 成立的条件下,$F\sim F(1,n-2)$,得到拒绝域:

$$W=\left\{F=\frac{U}{Q/(n-2)}\geqslant F_\alpha(1,n-2)\right\}$$

如果 $F>F_a$，则否定 H_0，即变量 y 与 x 之间存在线性相关关系；否则，接受 H_0，即变量 y 与 x 之间不存在线性相关关系. 此时可能有以下几种情况：

(1) y 对 x 没有显著影响，此时应该丢掉自变量 x.

(2) y 对 x 有显著影响，但这种情况不能用线性关系来表示，应该做非线性回归.

(3) 除 x 外还有其他不可忽略的变量对 y 也有显著影响，从而削弱了 x 对 y 的影响，此时对该问题的专业知识的了解往往起着重要作用.

3. 利用一元线性回归进行预测与控制

如何根据样本提供的信息来预测当变量 $x=x_0$ 时随机变量 Y_0 的值？一个自然的想法是：用预测量 $y_0=\beta_0+\beta_1 x_0$ 来代替，但是它与真值 Y_0 的差值是多少呢？预测量的优劣取决于 $|y_0-Y_0|$ 的大小，记为

$$d^2=1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}},\quad \hat{\delta}^2=\frac{Q}{n-2}$$

可以证明，当 Y_0 与 $Y_1,Y_2,\cdots,Y_n$ 相互独立时，$\dfrac{y_0-Y_0}{d\hat{\delta}}\sim t(n-2)$. 这样在显著性水平 α 下可得到 Y_0 的预测区间：

$$[y_0-t_a(n-2)d\hat{\delta},y_0+t_a(n-2)d\hat{\delta}]$$

当 n 较大时，预测区间的上下限近似取做 $y_0\pm 1.96\hat{\delta}$(可信度为 95%)或 $y_0\pm 2.58\hat{\delta}$(可信度为 99%).

控制是预测的反问题，即若要随机变量 Y 落在指定的区间 (y_L,y_U) 内，变量 x 应控制在什么区间内？从方程

$$\begin{cases}y_L=\beta_0+\beta_1 x_L-1.96\hat{\delta},\\ y_U=\beta_0+\beta_1 x_U-1.96\hat{\delta}\end{cases}$$

中解出 x_L 和 x_U，则当 $\beta_1>0$ 时，控制区间为 (x_L,x_U).

1.1.3 多元线性回归

在许多数学建模实际问题中，还会遇到一个随机变量与一组变量的相关关系问题，这要用到多元回归分析的方法来解决.

1. 多元线性回归的数学模型

设随机变量 $\boldsymbol{Y}$ 与 m 个变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m$ 有关系

$$\boldsymbol{Y}=\beta_0+\beta_1\boldsymbol{x}_1+\beta_2\boldsymbol{x}_2+\cdots+\beta_m\boldsymbol{x}_m+\boldsymbol{\varepsilon}$$

其中，$\boldsymbol{\varepsilon}$ 为随机项，且 $\varepsilon_i\sim N(0,\sigma^2)$. 记：

$$\boldsymbol{Y}=\begin{Bmatrix}y_1\\y_2\\\vdots\\y_n\end{Bmatrix},\quad \boldsymbol{X}=\begin{Bmatrix}1&x_{11}&\cdots&x_{1m}\\1&x_{21}&\cdots&x_{2m}\\\vdots&\vdots&&\vdots\\1&x_{n1}&\cdots&x_{nm}\end{Bmatrix}_{n\times(m+1)},\quad \boldsymbol{\varepsilon}=\begin{Bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{Bmatrix},\quad \boldsymbol{\beta}=\begin{Bmatrix}\beta_1\\\beta_2\\\vdots\\\beta_n\end{Bmatrix}$$

则随机变量 $\boldsymbol{Y}$ 与变量 $\boldsymbol{X}$ 的关系可化为 $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ 与一元线性回归情况类似，求回归系数 $\beta_0,\beta_1,\cdots,\beta_m$ 的最小二乘估计. 作残差平方和：

$$Q=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_{im}x_{im})^2$$

求 Q 分别关于 $\beta_0,\beta_1,\cdots,\beta_m$ 的一阶偏导数，并令它们等于零，得

$$\boldsymbol{\beta}=\begin{Bmatrix}\beta_1\\\beta_2\\\vdots\\\beta_n\end{Bmatrix}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

2. 多元线性回归的显著性检验

与一元回归情况类似，首先建立带检验假设：

$$H_0:\beta_1=\beta_2=\cdots=\beta_m=0$$

若能通过那个检验拒绝 H_0，则 $\boldsymbol{Y}$ 与 m 个变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m$ 之间存在线性相关关系.

记 $\bar{y}=\dfrac{1}{n}\sum_{i=1}^{n}y_i$，$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$，$U=S_{yy}-Q$.

选取检验统计量：

$$F=\frac{U/m}{Q/(n-m-1)}$$

在 H_0 成立的条件下，$F\sim F(m-1,n-m-1)$. 得到拒绝域：

$$W=\left\{F=\frac{U/m}{Q/(n-m-1)}\geqslant F_\alpha(m-1,n-m-1)\right\}$$

如果 $F>F_\alpha$，则否定 H_0，即 $\boldsymbol{Y}$ 与 m 个变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m$ 之间存在线性相关关系；否则，接受 H_0，即 $\boldsymbol{Y}$ 与 m 个变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m$ 之间不存在线性相关关系.

在多元线性回归模型中，拒绝假设 H_0，即回归方程显著. 然而变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m$ 对 $\boldsymbol{Y}$ 的影响并不都是十分重要的，人们还关心 $\boldsymbol{Y}$ 对 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m$ 的回归中哪些因素更重要些，哪些因素不重要. 要剔除不重要的，需要采用偏 F 检验

法，即检验假设：

$$H_k:\beta_k = 0, \quad k = 1,2,\cdots,m$$

通常选取统计量：

$$F_a = \frac{\beta_k^2 / a_{kk}}{Q/(n-m-1)}$$

其中，a_{kk}是矩阵$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$的主对角线上的第$k+1$个元素.

在H_k成立的条件下，$F_k \propto F(1,n-m-1)$. 得到拒绝域：

$$W = \left\{ F_k = \frac{\beta_k^2 / a_{kk}}{Q/(n-m-1)} \geqslant F_a(1,n-m-1) \right\}$$

如果$F_k > F_a$，拒绝H_k，即$\boldsymbol{x}_k$对$\boldsymbol{Y}$的影响显著；否则，接受H_k，即$\boldsymbol{x}_k$对$\boldsymbol{Y}$的影响不显著.

3. 预测问题

如何根据样本提供的信息来预测当变量$(x_1, x_2, \cdots, x_m) = (x_{01}, x_{02}, \cdots, x_{0m})$时随机变量$\boldsymbol{Y}_0$的值？一个自然的想法是用预测量：

$$\boldsymbol{y}_0 = \beta_0 + \beta_1 \boldsymbol{x}_{10} + \beta_2 \boldsymbol{x}_{20} + \cdots + \beta_m \boldsymbol{x}_{m0}$$

来替代. 预测量$\boldsymbol{y}_0$的优劣取决于$|\boldsymbol{y}_0 - \boldsymbol{Y}_0|$的大小. 记：

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad l_{ij} = \sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j), \quad i,j = 1,2,\cdots,m$$

$$\boldsymbol{L} = \begin{Bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & & \vdots \\ l_{m1} & \cdots & l_{mm} \end{Bmatrix}, \quad \boldsymbol{L}^{-1} = \begin{Bmatrix} l'_{11} & \cdots & l'_{1m} \\ \vdots & & \vdots \\ l'_{m1} & \cdots & l'_{mm} \end{Bmatrix}$$

$$d^2 = 1 + \frac{1}{n} + \sum_{i=1}^{m}\sum_{j=1}^{m} l'_{ij}(x_{0i} - \bar{x}_i)(x_{0j} - \bar{x}_j), \quad \hat{\delta}^2 = \frac{Q}{n-m-1}$$

可以证明当$\boldsymbol{Y}_0$与$\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n$相互独立时，$\dfrac{\boldsymbol{y}_0 - \boldsymbol{Y}_0}{d\hat{\delta}} \sim t(n-m-1)$. 这样在显著性水平$\alpha$下可得到$\boldsymbol{Y}_0$的预测区间：

$$[\boldsymbol{y}_0 - \boldsymbol{t}_a(n-m-1)d\hat{\delta}, \boldsymbol{y}_0 + \boldsymbol{t}_a(n-m-1)d\hat{\delta}]$$

上面介绍了一元回归方程和多元回归方程的理论基础，下面引入与线性模型有关的R函数.

适应于多元线性模型的基本函数是 lm()，其形式为

```
lm(formula, data, subset, weights, na.action, method = 'qr', model = TRUE,
x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ···)
```

其中 formula 为模型公式；data 为数据框；subset 为可选择向量，表示观察值的

子集；weights 为可选择向量，数据拟合的权重；返回值为线性模型结果的对象，存放在 fitted. model 中. 例如：

```
fm2<- lm(y~x1 + x2, data = production)
```

适用于 y 关于 x_1，x_2 的多元回归模型(隐含着截距项).

lm()函数的返回值称为拟合结果的对象，本质上是一个具有类属性值 lm 的列表 model，coeffcients，residuals 等成员. lm()的结果非常简单，为了获得更多的信息，可以使用对 lm()类对象有特殊操作的通用函数，这些函数包括：

```
addl  coef  effects  kappa  predict  residuals
alias  deviance  family  labels  print  step
anova  dropl  formula  plot  proj  summary
```

下面简单地介绍函数的使用方法.

(1) anova()函数. 其使用格式为

```
anova (object, …)
```

其中 object 是由 lm 或 glm 得到的对象，其返回值是模型的方差分析表.

(2) coefficients()函数简写形式为 coef()，其使用格式为

```
coefficients(object, …)
coef(object, …)
```

其中 object 是由模型构成的对象，其返回值是模型的系数.

(3) deviance()函数，其使用格式为

```
deviance(object, …)
```

其中 object 是由模型构成的对象，其返回值是模型的残差平方和.

(4) formula()函数. 其使用格式为

```
formula(object, …)
```

其中 object 是由模型构成的对象，其返回值是模型公式.

(5) plot()函数，其使用格式为

```
plot (object, …)
```

其中 object 是由 lm 构成的对象，绘制模型诊断的几种图形，显示残差、拟合值和一些诊断情况.

(6) predict()函数，其使用格式为

```
predict(object, newdata = data. frame)
```

其中 object 是由 lm 构成的对象. newdata 是预测点的数据，它由数据框形式输入，其返回值是预测值和预测区间.

(7) print()函数，其使用格式为

```
print(object, …)
```

其中 object 是由模型构成的对象，其返回值是显示模型拟合的结果，一般不用 print()而直接用键入对象的名称来显示.

(8) residuals()函数，其使用格式为

```
residuals (object, type = c ( " working"," response"," deviance"," pearson",
"partial"),
```

其中 object 是由 lm 或 aov 构成的对象. type 是返回值的类型，其返回值是模型的残差，简单的命令形式为 resid (object).

(9) step()函数，其使用格式为

```
step(object, …)
```

其中 object 是由 lm 或 glm 构成的对象，其返回值是逐步回归，根据 AIC (Akaike's Information Criterion)的最小值选择模型.

(10) summary()函数，其使用格式为

```
summary(object, …)
```

其中 object 是由 lm 构成的对象，其返回值是显示较为详细的模型拟合结果.

例 1.1 根据经验，在人的身高相等的情况下，血压的收缩压 Y 与体重(千克)、年龄(岁数)有关，现收集了 13 个男子的数据，见表 1.1. 试建立 Y 关于 X_1，X_2 的线性回归方程.

表 1.1 数据表

序号	X_1	X_2	Y	序号	X_1	X_2	Y
1	76.0	50	120	8	79.0	50	125
2	91.5	20	141	9	85.0	40	132
3	85.5	20	124	10	76.5	40	123
4	82.5	30	126	11	82.0	40	132
5	79.0	30	117	12	95.0	40	155
6	80.5	50	125	13	92.5	20	147
7	74.5	60	123				

解 R软件中的lm()同样可以求出回归系数,并作相应的检验.

下面是R软件的计算过程:

```
> blood <- data.frame(
    X1 = c(76.0, 91.5, 85.5, 82.5, 79.0, 80.5, 74.5, 79.0, 85.0, 76.5, 82.0,
95.0, 92.5),
    X2 = c(50, 20, 20, 30, 30, 50, 60, 50, 40, 55, 40,40,20),
    Y = c(120, 141, 124, 126, 117, 125, 123, 125, 132, 123, 132, 155, 147)
    )
> lm.sol <- lm(Y ~ X1 + X2, data = blood)
> summary(lm.sol)
Call:
lm(formula = Y ~ X1 + X2, data = blood)
Residuals :
    Min      1Q  Median      3Q     Max
-4.0404 -1.0183  0.4640  0.6908  4.3274
Coefficients :
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -62.96336   16.99976  -3.704 0.004083 **
X1            2.13656    0.17534  12.185 2.53e-07 ***
X2            0.40022    0.08321   4.810 0.000713 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05' .' 0.1 ' ' 1
Residual standard error: 2.854 on 10 degrees of freedom
Multiple R-Squared: 0.9461, Adjusted R-squared: 0.9354
F-statistic: 87.84 on 2 and 10 DF, p-value: 4.531e-07
```

从计算结果可以得到,回归系数与回归方程的检验都是显著的,因此,回归方程为

$$\hat{Y} = -62.96 + 2.136X_1 + 0.4002X_2$$

```
>  source("beta.int.R")
>  beta.int(lm.sol)
                Estimate        Left       Right
(Intercept) -62.9633591 -100.8411862 -25.0855320
x1            2.1365581    1.7458709   2.5272454
x2            0.4002162    0.2148077   0.5856246
```

其中R程序beta.int.R描述的是参数向量β的置信水平为$1-\alpha$的区间估计程序:

```
beta.int <- function (fm, alpha = 0.05){
```

```
A<- summary(fm) $ coefficients
        df<- fm $ df .residual
        left<- A[,i] - A[,2] * qt(1 - alpha/2, df)
        right<- A[,1] + A[,2] * qt(1 - alpha/2, df)
        rowname<- dimnames (A)[ [1]]
        colname<- c("Estimate", "Left", "Right")
        matrix(c(A[,1], left, right), nc01 = 3, dimnames = list(rowname,
colname))
    }
```

其中 summary 是提取模型信息，返回值为一列表，其中 $ coefficients 是由回归系数、标准差、t 值和 P 值构成的矩阵，若 fm 是由 Im 计算得到回归模型，其中 $ df. residual 为模型的自由度.

1.1.4 曲线回归模型

在许多数学建模实际问题中，一个随机变量与另一个随机变量的关系不是线性关系，而是某种曲线关系. 那么如何确定回归方程呢？常用的有 3 种方法，即：化为一元线性回归、多项式回归和分段回归.

1. 化为一元线性回归

在某些非线性回归方程中，为了确定其中的未知参数，一般将非线性回归方程转化为线性回归方程，通过线性回归方程的参数估计将非线性回归方程的参数估计出来. 表 1.2 列出了常用的可线性化回归方程($a>0$).

表 1.2 常用的可线性化回归方程

曲线方程	变换公式	变换后的线性方程
$1/y=a+b/x$	$u=1/x, v=1/y$	$v=a+bu$
$y=ax^b$	$u=\ln x, v=\ln y$	$v=c+bu(c=\ln a)$
$y=a+b\ln x$	$u=\ln x, v=y$	$v=a+bu$
$y=ae^{bx}$	$u=x, v=\ln y$	$v=c+bu(c=\ln a)$
$y=ae^{b/x}$	$u=1/x, v=\ln y$	$v=c+bu(c=\ln a)$
$y=1/(a+be^{-x})$	$u=e^{-x}, v=1/y$	$v=a+bu$

2. 多项式回归

在曲线回归中，比较困难的是选择合适的曲线类型. 有的曲线也不一定经变

化就能化为直线形状.这就引出了解决曲线回归的另一种方法——一元多项式回归分析方法,即回归函数 $y=f(x)$ 是一个多项式:

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_m x^m$$

其中 $m\geqslant 2$.随机变量 Y 与 x 之间的相关关系为

$$Y=\beta_0+\beta_1 x_1+\beta_2 x^2+\cdots+\beta_m x^m+\varepsilon$$

其中,ε 为随机项,且 $\varepsilon_i\sim N(0,\sigma^2)$.对自变量 x 作变换,

$$x_j=x^j,\quad j=1,2,\cdots,m$$

得到

$$Y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m+\varepsilon$$

再将原来的多项式回归问题中的 n 对数据 $(x_i,y_i)(i=1,2,\cdots,n)$ 相应地变换成:

$$(y_i;x_{i1},x_{i2},\cdots,x_{im}),\quad i=1,2,\cdots,n$$

其中,$x_{ij}=x_i^j$,$j=1,2,\cdots,n$,$j=1,2,\cdots,m$.

这样我们便能使用多元线性回归分析的方法进行处理.

3. 分段回归

当散点图上呈现的趋势较为复杂时,常常难以找到某种合适的变换为线性回归.用多项式回归时也不能得到令人满意的效果,即使用高达 20 次的多项式也拟合得不太好,这表现在残差平方和并不随着多项式次数的增加而迅速下降.究其原因,这是因为自变量 x 变化范围较大,而在 x 不同的范围内,y 的相应规律可能并不相同.为了解决这个问题,可考虑进行分段回归.

分段回归有下面几个步骤和要求:

(1) 将自变量范围分成若干段,选择好适合的分点.

(2) 在每段内用一个次数不高(通常不超过 3 次)的多项式来拟合.

(3) 在两段接头之处,曲线相连而且要连接得光滑.

例 1.2 某大型牙膏制造企业为了更好地拓展产品市场,有效地管理库存,公司董事会要求销售部门根据市场调查,找出公司生产的牙膏销售量与销售价格、广告投入等之间的关系,从而预测出在不同价格和广告费用下的销售量,为此,销售部的研究人员收集了过去 30 个销售周期(每个销售周期为 4 周)公司生产的牙膏的销售量、销售价格、投入的广告费用,以及周期内其他厂家生产同类牙膏的市场平均销售价格,如表 1.3 所示,试根据这些数据建立一个数学模型,分析牙膏销售量与其他因素的关系,为制定价格策略和广告投入策略提供数量依据.

表 1.3 牙膏销售量与销售价格、广告费用等数据

销售周期	公司销售价格/元	其他厂家平均价格/元	价格差/元	广告费用/百万元	销售量/百万支
1	3.85	3.80	−0.05	5.50	7.38
2	3.75	4.00	0.25	6.75	8.51
3	3.70	4.30	0.60	7.25	9.52
4	3.70	3.70	0.00	5.50	7.50
5	3.60	3.85	0.25	7.00	9.33
6	3.60	3.80	0.20	6.50	8.28
7	3.60	3.75	0.15	6.75	8.75
8	3.80	3.85	0.05	5.25	7.87
9	3.80	3.65	−0.15	5.25	7.10
10	3.85	4.00	0.15	6.00	8.00
11	3.90	4.10	0.20	6.50	7.89
12	3.90	4.00	0.10	6.25	8.15
13	3.70	4.10	0.40	7.00	9.10
14	3.75	4.20	0.45	6.90	8.86
15	3.75	4.10	0.35	6.80	8.90
16	3.80	4.10	0.30	6.80	8.87
17	3.70	4.20	0.50	7.10	9.26
18	3.80	4.30	0.50	7.00	9.00
19	3.70	4.10	0.40	6.80	8.75
20	3.80	3.75	−0.05	6.50	7.95
21	3.80	3.75	−0.05	6.25	7.65
22	3.75	3.65	−0.10	6.00	7.27
23	3.70	3.90	0.20	6.50	8.00
24	3.55	3.65	0.10	7.00	8.50
25	3.60	4.10	0.50	6.80	8.75
26	3.65	4.25	0.60	6.80	9.21
27	3.70	3.65	−0.05	6.50	8.27
28	3.75	3.75	0.00	5.75	7.67
29	3.80	3.85	0.05	5.80	7.93
30	3.70	4.25	0.55	6.80	9.26

记牙膏销售量为 Y，价格差为 X_1，公司的广告费为 X_2，假设基本模型为线性模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

输入数据，调用 R 软件中的 lm() 函数求解，并用 summary() 显示计算结果(程

序名：exam1206.R).

```
> toothpaste<-data.frame(
    X1=c(-0.05, 0.25,0.60,0, 0.25,0.20,0.15, 0.05,-0.15,0.15,
         0.20,0.10,0.40,0.45,0.35,0.30,0.50,0.50, 0.40,-0.05,
         -0.05, -0.10,0.20, 0.10,0.50,0.60, -0.05,0,0.05,0.55),
    X2=c(5.50,6.75,7.25,5.50,7.00, 6.50,6.75,5.25, 5.25,6.00,
         6.50,6.25,7.00,6.90, 6.80, 6.80,6.50,5.75, 5.80, 6.80),
    Y=c(7.38,8.51,9.52,7.50, 9.33, 8.28,8.75,7.87,7.10,8.00,
        7.89,8.15, 9.10, 8.86,8.90, 8.87,9.26,9.00,8.75,7.95,
        7.65,7.27,8.00,8.50,8.75, 9.21, 8.27,7.67,7.93, 9.26)
    )
> lm.sol<- lm(Y~X1+X2, data=toothpaste)
> summary(lm.sol)
Call:
lm(formula = Y ~ X1 + X2)
Residuals :
     Min         1Q          Median       3Q         Max
-0.497785  -0.120312  -0.008672  0.110844    0.581059
Coefficients :
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.4075   0.7223    6.102   1.62e-06 ***
X1          1.5883   0.2994    5.304   1.35e-05 ***
X2          0.5635   0.119     4.733   6.25e-05 ***
---
Signif. codes: 0 . *** . 0.001 . ** , 0.01, * . 0.05 ... 0.1 . , 1
Residual standard error: 0.2383 0n 27 degrees of freedom
Multiple R-Squared: 0.886, Adjusted R-squared: 0.8776
F-statistic: 105 0n 2 and 27 DF, p-value: 1.845e-13
```

计算结果通过回归系数检验和回归方程检验，由此得到销售量与价格差与广告费之间的关系为

$$Y = 4.4075 + 1.5883X_1 + 0.5635X_2$$

模型的进一步分析：我们画出 Y 与 X_1 和 Y 与 X_2 散点图，从散点图上可以看出，对于 Y 与 X_1，用直线拟合较好，而对于 Y 与 X_2，则用二次曲线拟合较好，如图 1.1 所示.

绘 Y 与 X_1 的散点图和回归直线：

```
> attach(toothpaste)
> plot(Y~X1); abline(lm(Y~X1))
```

绘 Y 与 X_2 的散点图和回归曲线：

```
> lm2.sol<-lm(Y~X2+I(X2^2))
```

```
> x<- seq(min(X2), max(X2), len = 200)
> y<- predict (lm2.sol, data.frame(X2 = x))
> plot(Y~X2); lines(x,y)
```

其中 I(X2 ^2)表示模型中 X_2 的平方项，即 X_2^2.

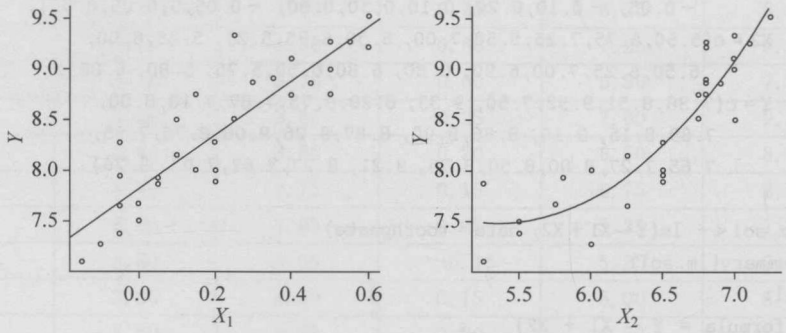

图 1.1 X_1，X_2 和散点图及拟合曲线

从图 1.1 看出，将销售量模型改为

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \varepsilon$$

似乎更合理，我们作相应的回归分析：

```
> lm.new<- update(lm.sol, .~. + I(X2 ^2))
> summary(lm.new)
Call:
lm(formula = Y~Xl + X2 + I(X2 ^2), data = toothpaste)
Residuals :
     Min       1Q      Median      3Q      Max
-0.40330 -0.14509 -0.03035   0.15488   0.46602
Coefficients :
            Estimate Std. Error t value Pr(>ltl)
(Intercept) 17.3244   5.6415    3.071  0.004951 **
Xl           1.3070   0.3036    4.305  0.000210 ***
X2          -3.6956   1.8503   -1.997  0.056355 .
I(X2 ^2)     0.3486   0.1512    2.306  0.029341 *
-- -
Signif. codes: 0 . *** . 0.001 . ** , 0.01, * . 0.05 ... 0.1 ', 1
Residual standard error: 0.2213 0n 26 degrees of freedom
Multiple R-Squared: 0.9054, Adjusted R-squared: 0.8945
F-statistic: 82.94 0n 3 and 26 DF, p-value: 1.944e-13
```

此时，我们发现，模型残差的标准差 $\hat{\sigma}$ 有所下降，相关系数的平方 R^2 有所上升，这说明模型修正是合理的，但这也出现一个问题，就是对应于 β_2 的 $P>$

0.05. 为进一步分析，作 β 的区间估计.

```
> source("beta.int.R")
> beta.int(lm.new)
                Estimate        Left         Right
(Intercept)  17.3243685   5.72818421   28.9205529
X1            1.3069887   0.68290927    1.9310682
X2           -3.6955867  -7.49886317    0.1076898
I(X2 2)       0.3486117   0.03786354    0.6593598
```

β_2 的区间估计是[−7.49886317, 0.1076898]，它包含了 0，也就是说，β_2 的值可能会为 0. 去掉 X_2 的一次项，再进行分析：

```
> lm2.new<-update(lm.new, .~. -X2)
> summary(lm2.new)
Call:
lm(formula = Y~X1 + I(X2^2), data = toothpaste)
Residuals:
     Min       1Q   Median       3Q      Max
-0.485943 -0.114094 -0.004604 0.105342 0.559195
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.07667    0.35531   17.102  5.17e-16 ***
X1          1.52498    0.29859    5.107  2.28e-05 ***
I(X2^2)     0.04720    0.00952    4.958  3.41e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2332 on 27 degrees of freedom
Multiple R-Squared: 0.8909, Adjusted R-squared: 0.8828
F-statistic: 110.2 on 2 and 27 DF, p-value: 1.028e-13
```

此模型虽然通过了 F 检验和 t 检验，但与上一模型对比来看，$\hat{\sigma}$ 上升，R^2 下降. 这又是此模型的不足之处.

再作进一步的修正，考虑 X_1, X_2 交互作用，即模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \beta_4 X_1 X_2 + \varepsilon$$

```
> lm3.new<-update(lm.new, .~ . + X1 * X2)
> summary(lm3.new)
Call:
lm(formula = Y ~ X1 + X2 + I(X2-2) + X1:X2, data = toothpaste)
Residuals:
     Min       1Q   Median       3Q      Max
-0.437250 -0.117540 0.004895 0.122634 0.384097
```

```
Coefficients :
                Estimate  Std. Error t  value  Pr(> ltl)
(Intercept)     29.1133   7.4832        3.890  0.000656 ***
Xl              11.1342   4.4459        2.504  0.019153 *
X2              -7.6080   2.4691       -3.081  0.004963 **
I(X2 - 2)        0.6712   0.2027        3.312  0.002824 **
X1:X2           -1.4777   0.6672       -2.215  0.036105 *
-- -
Signif. codes: 0 '*** '0.001 '** .'0.01 '* '0.05 '.' 0.1 ' '1
Residual standard error: 0.2063 0n 25 degrees of freedom
Multiple R - Squared: 0.9209, Adjusted R - squared: 0.9083
F - statistic: 72.78 0n 4 and 25 DF, p - value: 2.107e - 13
```

模型通过 t 检验和 F 检验，并且 $\hat{\sigma}$ 减少，R^2 增加，因此，最终模型选为

$$Y = 29.1133 + 11.1342X_1 - 7.6080X_2 + 0.6712X_2^2 - 1.4777X_1X_2 + \varepsilon$$

1.2 非线性规划模型

非线性规划是计算数学和运筹学交叉的学科，对非线性规划模型的研究源于实际生活中对问题进行更为精确的描述并解决的迫切需要. 非线性规划理论是基于线性规划理论发展起来的，自 20 世纪 40 年代人们获得求解线性规划问题的单纯形法之后，线性规划在理论上日趋成熟，并在实践中获得广泛的应用，然而随着社会各个方面的发展，许多实际问题用线性规划理论建立模型并解决的效果并不好，这种情形促使科学研究转移到非线性领域. 近几十年来许多专家和学者为有效地求解非线性规划问题付出了艰辛的探索，促使非线性规划理论不断成熟和完善. 对非线性规划算法的研究主要集中在如何提高算法的效率以及扩大算法的适用范围. 按数学空间划分，理论上可将非线性理论的研究分为有穷维优化和无穷维优化，由于现实条件的限制，目前已经实现的大多数非线性规划算法只限于有限维空间，而且相应的理论也比较成熟. 在无穷维空间上，随着许多优化学者和专家的深入研究，这方面的研究成果不断涌现，将很多有限维空间的理论推广到无穷维空间.

1.2.1 非线性规划算法综述

非线性规划模型按照目标函数的多少，可分为单目标规划和多目标规划. 这两类规划模型在算法上既有联系又有区别，下面分别作简要介绍.

1. 单目标规划算法

求解单目标非线性规划模型的算法大都属于迭代算法，其中只有一个变量的优化问题的计算方法是优化方法中最基本的，是多变量问题求解中迭代过程的重要步骤. 而对于多维优化模型，目前在算法上可以分为两类：一类是线搜索(line search)方法；另一类是信赖域(trust region)方法.

目前非线性规划算法的研究主要集中在线搜索方法. 线搜索方法有两个重要的环节：一是搜索方向的选取，二是迭代步长的确定，这两个环节的差异衍生出许多不同的算法. 搜索方向的产生依赖于子问题的构造，由此衍生出的方法有梯度法、共轭梯度法、拟牛顿法、变尺度法等.

信赖域法是一类较新的方法，自 20 世纪 80 年代以来有很多文章对这一领域进行了研究，目前虽然没有线搜索法成熟，但由于其本身具有较强的收敛性和可靠性，这一问题仍吸引了很多优化专家和学者.

2. 多目标规划算法

在实际生产与分配的活动中有许多问题，它们具有多个彼此联系而又矛盾的决策目标，这些问题的出现促使了多目标最优化问题的产生. 多目标规划问题成为正式学科始于 20 世纪 50 年代，由最初研究数量经济问题到关于向量极值问题的研究，为多目标最优化问题的研究奠定了良好的基础. 目前多目标最优化问题不仅在理论上逐渐成熟，而且在工程技术、经济、管理及系统工程等许多领域获得了广泛的应用，引起了许多学者和专业人员的重视.

多目标最优化问题是在单目标规划问题的基础上建立起来的，因而主要是研究如何建立多目标与单目标问题的联系和如何衡量解的优劣，以及存在性、稳定性等. 常见的多目标规划算法有主要目标法、线性加权和法、理想点法和评价函数法等，下面介绍多目标优化中一个常用的算法——线性加权和法.

线性加权和法是根据 p 个单目标函数 $f_j(x)(j=1,2,\cdots,p)$ 的重要程度，给以不同的权重 $\lambda_j(j=1,2,\cdots,p)$，然后简单相加构成单目标优化问题的目标函数在多目标优化问题的约束集合 $\mathbf{R}$ 上求最优解. 所构造的单目标问题形式如下：

$$\min_{x\in\mathbf{R}} u(x)\sum_{j=1}^{p}\lambda_j f_j(x)=\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{f}(x)$$

称此单目标问题的解叫做原多目标问题在线性加权和意义下的最优解. 这里

$$\boldsymbol{f}(x)=(f_1(x),f_2(x),\cdots,f_p(x))^{\mathrm{T}},\quad \boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_p)^{\mathrm{T}}$$

其中 $\wedge^{+}=\left\{\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_p)^{\mathrm{T}}\mid \lambda_j\geqslant 0,\sum_{j=1}^{p}\lambda_j=1\right\}$，$\boldsymbol{\lambda}$ 称为权向量.