第5章 线性回归分析的R语言应用

在现实世界中，一个事物的运动变化总是与其他事物相关联，有的存在因果关系。这种因果关系有的是线性的，有的是非线性的。当预测对象与其影响的关系是线性的，且只有一个影响因素时，则可用一元线性回归方法建立其一元线性回归预测模型，来表述和分析其因果关系；当有两个或多个影响因素同时作用于一个预测对象时，则用多元线性回归法建立多元线性回归预测模型。

5.1 一元线性回归分析基本理论

5.1.1 一元线性回归分析的OLS估计

一元线性回归预测模型是最基本的回归模型，其数学表达式是

$$\hat{y} = \beta_0 + \beta_1 x + \varepsilon \tag{5-1}$$

式中：$\hat{y}$——预测对象，因变量或被解释变量的预测值；

x——影响因素，自变量或解释变量的相应值；

β_0、β_1——待估计的参数，称为回归系数；

ε——偏差，或估计误差，或残差，有些书用 e 表示。

为了估计参数 β_0、β_1，最常用的方法是最小二乘法。首先，要收集预测对象 y 及相关因素 x 的数据样本 n 对（实际值）：

$$(y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n)$$

再将其描绘在坐标图上（x 为横轴，y 为纵轴），当这 n 对数据点近似呈直线分布时，则可以用一元线性回归模型（见式(5-1)），式中 $\beta_0 + \beta_1 x = y$ 应是预测对象的实际值，因而对应样本中的每一个 x_i 都有一个 y_i 的估计值 $\hat{y}_i$，$i=1,2,\cdots,n$；y_i 与 $\hat{y}_i$ 之间存在一个偏差 ε_i，于是有

$$\varepsilon_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$$

设

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

可见，Q 是参数 β_0、β_1 的函数。为了求 Q 的最小值，可利用极值原理：

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

即

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

求解此联立方程可得

$$\beta_1 = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}, \quad \beta_0 = \frac{1}{n}\sum_{i=1}^{n}y_i - \frac{\beta_1}{n}\sum_{i=1}^{n}x_i$$

令

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

简记 $\sum_{i=1}^{n}$ 为 $\sum$，则有

$$\beta_1 = \frac{\sum x_i y_i - \bar{x}\sum y_i}{\sum x_i^2 - \bar{x}\sum x_i}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

例 5-1：某地某工业部门近 8 年来专门人才数 y（百人）与职工人数 x（万人）的数据如表 5-1 所示。

表 5-1　近 8 年来专门人才数（百人）与职工人数（万人）数据

序号	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1.30	4.88	1.6900	23.8144	6.3440
2	1.34	5.19	1.7956	26.9361	6.9546
3	1.40	6.74	1.9600	45.4276	9.4360
4	1.41	7.31	1.9881	53.4361	10.3071
5	1.42	8.23	2.0164	67.7329	11.6866
6	1.53	10.41	2.3409	108.3681	15.9273
7	1.55	11.10	2.4025	123.2100	17.2050
8	1.60	11.80	2.5600	139.2400	18.8800
合计	11.55	65.66	16.7537	588.1652	96.7406

按表 5-1 中的数据可算得，$\bar{x}=1.4438$，$\bar{y}=8.2075$。

$$\beta_1 = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\bar{x}^2} = \frac{96.7406 - 8\times 1.4438\times 8.2075}{16.7535 - 8\times 1.4438^2} = 24.863$$

$$\beta_0 = \bar{y} - \beta_1\bar{x} = 8.2075 - 24.863\times 1.4438 = -27.688$$

则得一元线性回归模型为：

$$\hat{y} = -27.688 + 24.863x \tag{5-2}$$

5.1.2　一元线性回归模型的统计检验

回归模型建立以后，它与实际数据拟合如何？模型的线性关系显著性如何？模型的

有效性如何？要解答这些问题，必须进行数理统计和经济意义的检验。常规的统计检验如下。

1. 标准离差检验

一般用标准离差 s 来检验回归模型预测的精度，算式为

$$s=\sqrt{\frac{1}{n-2}\sum(y_i-\hat{y}_i)^2}$$

希望 s 值愈小愈好，一般要求 $s/\bar{y}<10\%$，根据实际情况可适当放宽到 15%。

对表 5-1 所示的数据，可算得 $s/\bar{y}=0.0479=4.79\%<10\%$，故可认为所得模型式(5-2)有较好的精度。

2. 相关系数检验

可以用相关系数 r 来检验 y 与 x 两变量之间的线性相关的显著程度，其算式为

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \tag{5-3}$$

在数学上可以证明：$-1\leqslant r\leqslant 1$，即 $|r|\leqslant 1$，于是有：

- 当 $|r|=1$ 时，实际 y_i 完全落在回归直线上，y 与 x 完全线性相关。
- 当 $0<r<1$ 时，y 与 x 有一定的正线性相关，愈接近 1 则愈好。
- 当 $-1<r<0$ 时，y 与 x 有一定的负线性相关，愈接近 -1 则愈好。

实际的检验操作方法如下：

(1) 按式(5-3)算出相关系数 r 的值。

(2) 拟定显著性水平 α(一般取 $\alpha=0.05$，即 95%的置信度)，再查相关系数表，查表时取自由度 $v=n-2$，得相关系数临界值 r_α。

(3) 判别如下：

- 当 $|r|\geqslant r_\alpha$ 时，y 与 x 在 α 显著水平下显著相关，检验通过。
- 当 $|r|<r_\alpha$ 时，y 与 x 的线性关系不显著，检验未通过。

也可以通过如下途径解决：

要对相关系数的显著性进行检验，首先提出原假设 $H_0:\rho=0$(总体相关系数为 0，表示总体的两变量之间线性相关性不显著)，备择假设 $H_1:\rho\neq 0$(总体相关系数不为 0，表示总体的两变量之间线性相关性显著)。可以证明，当原假设 $H_0:\rho=0$ 成立时，统计量 t 是服从自由度为 $n-2$ 的 t 分布，即

$$t=r\sqrt{n-2}/\sqrt{1-r^2}\sim t(n-2)$$

对于给定的显著性水平 α，查 t 分布表得临界值 $t_{\alpha/2}(n-2)$，将 t 值与临界值进行比较：

- 当 $|t|<t_{\alpha/2}(n-2)$，接受 H_0，表示总体的两变量之间线性相关性不显著。
- 当 $|t|\geqslant t_{\alpha/2}(n-2)$，拒绝 H_0，表示总体的两变量之间线性相关性显著(即样本相关系数的绝对值接近 1，并不是由于偶然机会所致)。

以如表 5-2 所示的数据为例，检验能源消耗量与工业总产值之间的线性相关性是否显著($\alpha=0.05$)。

表 5-2　某地能源消耗量与工业总产值的相关表

能源消耗量(十万吨)	工业总产值(亿元)	能源消耗量(十万吨)	工业总产值(亿元)
35	24	62	41
38	25	64	40
40	24	65	47
42	28	68	50
49	32	69	49
52	31	71	51
54	37	72	48
59	40	76	58

由表 5-2 的数据计算出的相关系数为

$$r=\frac{S_{xy}}{S_xS_y}=\frac{\sum(x-\bar{x})(y-\bar{y})/n}{\sqrt{\sum(x-\bar{x})^2/n}\sqrt{\sum(y-\bar{y})^2/n}}$$

$$=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-\left(\sum y\right)^2}}=0.9757$$

式中，S_{xy}是变量 x、y 的样本协方差，S_x、S_y 分别为变量 x、y 的样本标准差。

提出原假设和备择假设：

$$H_0:\rho=0 \quad H_1:\rho\neq 0$$

当 $H_0:\rho=0$ 成立时，则统计量为

$$t=r\sqrt{n-2}/\sqrt{1-r^2}\sim t(n-2)$$

实际计算如下：

$$t=0.9757\sqrt{16-2}/\sqrt{1-(0.9757)^2}=16.6616$$

对于给定的 α，查 t 分布表得临界值：

$$t_{\alpha/2}(n-2)=t_{0.025}(14)=2.1448$$

$$|t|=16.6616>t_{0.025}(14)=2.1448$$

所以拒绝原假设，表示总体的两变量之间线性相关性显著，即说明能源消耗量与工业总产值之间存在显著的线性相关关系，所拟合的线性回归方程具有 95%的置信概率。

3. *F* 检验

F 检验用来检验 y 与 x 之间是否存在显著的线性统计关系。*F* 检验值用下式计算：

$$F=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)^2}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)/(n-2)} \tag{5-4}$$

或

$$F=\frac{r^2}{(1-r^2)/(n-2)} \tag{5-5}$$

检验操作方法如下：

(1) 按式(5-4)、式(5-5)算出 F 的值。

(2) 拟定显著性水平 α(一般取 $\alpha=0.05$,即 95%的置信度),取自由度 $v=n-2$,查 F 检验表,得 F 临界值 F_α。

(3) 判别如下:

- 当 $F>F_\alpha$时,y 与 x 在 α 显著水平下存在线性统计关系,检验通过,所建模型有效。
- 当 $F<F_\alpha$时,检验未通过,所建模型无效。

5.1.3 一元线性回归模型预测的置信区间

一元线性回归模型经过以上检验通过后可用于预测,一般将各项检验值 r、F、DW 注在回归模型之下,以示通过检验的结果。预测时,可将新的自变量 x_0(例如计划值或者由其他模型获得的数值)代入回归计算得出相应的预测值 y_0,但由于预测值有一定的误差,亦即预测结果有一定的波动范围,此范围称为置信区间,其计算方法如下:

(1) 按 $S=\sqrt{\frac{1}{n-2}\sum(y_i-\hat{y}_i)^2}$ 算出标准差 S 的值。

(2) 求算置信区间。

当样本量较大($n\geqslant30$),并取置信度为 $100\times(1-\alpha)$时,则置信区间为

$$\hat{y}\pm t_{\alpha/2}S$$

式中,$t_{\alpha/2}$——显著性水平 α,自由度 $n-2$ 时的 t 统计量,可查 t 检验表取得。

当样本量比较小时($n<30$),首先应对标准差进行修正,其修正系数 c_0 按下式计算:

$$c_0=\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_0-\bar{x})^2}}$$

然后计算置信区间:

$$\hat{y}\pm t_{\alpha/2}c_0S$$

5.2 一元线性回归分析的 R 语言应用

例 5-2:某公司为研究销售人员数量对新产品销售额的影响,从其下属多家公司中随机抽取 10 个子公司,这 10 个子公司当年新产品销售额和销售人员数量统计数据如表 5-3 所示。试用简单回归分析方法研究销售人员数量对新产品销售额的影响。

表 5-3 新产品销售额和销售人员数量统计数据

地区	新产品销售额/万元	销售人员数量/人
1	385	17
2	251	10
3	701	44
4	479	30
5	433	22
6	411	15
7	355	11
8	217	5
9	581	31
10	653	36

在目录G:\2glkx\data下建立al5-1.xls数据文件后,使用的命令如下:

```
> library(RODBC)  #使用此命令时必须先安装RODBC,见3.9.2节
> z<-odbcConnectExcel("G:/2glkx/data/al5-1.xls")
> sq<-sqlFetch(z,"Sheet1")
> close(z)
> sq
```

执行以上5行命令后,得到如下结果:

```
     dq  xse  rs
1     1   385  17
⋮
10   10   653  36
```

1. 对数据进行描述性分析

在符号“>”后输入如下命令:

```
> y<-sq$xse; x<-sq$rs
> d<-data.frame(y,x)
> summary(d)  #这些命令是对年份、通货膨胀率、失业率等变量进行详细描述性分析
```

输入以上3行命令后,按回车键,得到如下分析结果。

```
       y          x
Min.   :217.0   Min.   : 5.00
1st Qu.:362.5   1st Qu.:12.00
Median :422.0   Median :19.50
Mean   :446.6   Mean   :22.10
3rd Qu.:555.5   3rd Qu.:30.75
Max.   :701.0   Max.   :44.00
```

通过观察上面的结果,可以得到很多信息,包括2个最小值、2个第一百分位数、2个中位数、2个平均值、2个最大值等。信息描述如下。

(1) 最小值(Smallest)。

变量xse的最小值是217.0。

变量rs的最小值是5.00。

(2) 百分位数。

可以看出,变量xse的第1个四分位数(25%)是362.5,第3个四分位数(75%)是555.5。变量rs的第1个四分位数(25%)是12.00,第3个四分位数(75%)是30.75。

(3) 中位数(median)。

变量xse的中位数是422.0。

变量unwork的中位数是19.50。

(4) 平均值(Mean)。

变量xse的平均值是446.6。

变量rs的平均值是22.10。

(5) 最大值(Largest)。

变量 xse 的最大值是 701.0。

变量 rs 的最大值是 44.00。

2. 对数据进行相关分析

在符号“>”后输入如下命令:

```
> cor(y,x)   # 本命令是对新产品销售额、销售人员人数等变量进行相关性分析
```

输入完后,按回车键,得到如下分析结果:

```
[1]  0.9699062
```

通过观察上面的结果,可以看出 xse 和 rs 之间的相关系数为 0.9699062,这说明两个变量之间存在很强的正相关关系,所以可以对其进行回归分析。

3. 对数据进行回归分析

在符号“>”后输入如下命令:

```
> lm.reg <- lm(y~1 + x)   # 本命令是对 xse、rs 等变量进行简单回归分析
> summary(lm.reg)
```

每输入完一行命令后,按回车键,最后得到如下分析结果:

```
Call:
lm(formula = y ~ 1 + x)
Residuals:
    Min      1Q  Median      3Q     Max
-64.225 -18.702  -5.799  33.678  51.240
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.295   27.327    6.451  0.000198 ***
x            12.231    1.086   11.267  3.46e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 41.38 on 8 degrees of freedom
Multiple R-squared: 0.9407, Adjusted R-squared: 0.9333
F-statistic: 126.9 on 1 and 8 DF, p-value: 3.46e-06
```

通过观察上面的结果,可以看出模型的 F 值=126.9,p 值为 0,说明该模型整体上是非常显著的。模型的可决系数 R-squared=0.9407,修正的可决系数 Adjusted R-squared=0.9333,说明模型的解释能力是很强的。

模型的回归方程是

$$xse = 12.231 \times rs + 176.295$$ (其中 xse 表示新产品的销售额)

变量 rs 的系数标准误是 1.086,t 值为 11.267,p 值为 0.00,系数是非常显著的。常数项的系数标准误是 27.327,t 值为 6.451,p 值为 0.000198,变量 x 的系数是非常显著的。

4. 求参数的置信区间

R 语言软件可以用函数 confint()求参数的置信区间。

```
> confint(lm.reg,level = 0.95)
```

执行以上命令后，得到如下结果：

```
                2.5 %    97.5 %
(Intercept)  113.279335  239.31107
x              9.727714   14.73426
```

5. 预测分析

若要求 rs＝40 时相应 xse 的置信水平为 0.95 的预测值和预测区间，可用 predict()函数求预测值和预测区间。

```
> point <- data.frame(x = 40)
> lm.pred <- predict(lm.reg,point,interval = "prediction",level = 0.95)
> lm.pred
```

执行以上 3 行命令后，得到如下结果：

```
       fit      lwr      upr
1 665.5347 555.8868 775.1825
> q()       #退出 R 语言
```

注意：

(1) 对线性的第一种解释是指 y 是 x 的线性函数，比如，$y=\alpha+\beta x$。

(2) 对线性的第二种解释是指 y 是参数的一个线性函数，它可以不是变量 x 的线性函数。比如，$y=\alpha+\beta x^2$ 就是一个线性回归模型，但 $y=\alpha+\sqrt{\beta}x$ 则不是。

(3) 这里线性回归一词总是指参数 β 为线性的一种回归(即参数只以一次方出现)，对解释变量 x 则可以是或不是线性的。

(4) 有些模型看起来不是线性回归模型，但经过一些基本代数变换可以转换成线性回归模型。例如，$y_t=Ax_t^{\beta}e^{ut}$，通过做对数变换，可以转化为线性回归模型。

5.3　多元线性回归分析基本理论

5.3.1　多元线性回归模型假设

当预测对象 y 同时受到多个解释变量 $x_1,x_2,\cdots,x_m$ 影响，且各个 $x_j(j=1,2,\cdots,m)$ 与 y 都近似地表现为线性相关时，则可建立多元线性回归模型来进行预测和分析，模型为

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_mx_m+\varepsilon_i \tag{5-6}$$

对 $i(i=1,2,\cdots,n)$ 个样本均可写出

$$y=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\cdots+\beta_mx_{mi}+\varepsilon_i,\quad i=1,2,\cdots,n \tag{5-7}$$

式中，$\beta_0,\beta_1,\cdots,\beta_m$——模型的回归系数；

ε——随机干扰误差。

模型(5-6)可以用最小二乘法来估计其参数，也可用矩阵解法。

此模型的基本假定如下：

假定 1：解释变量 X 不是随机变量。在一个样本中，X 的值不能完全相同。

假定 2：误差项的均值为 0，即 $E(\varepsilon_i|X_i)=0$。

假定 3：误差项同方差，即 $\mathrm{Var}(\varepsilon_i|X_i)=\sigma^2$，$i=1,2,\cdots,n$。

假定 4：误差项无序列相关，即 $\mathrm{Cov}(\varepsilon_i,\varepsilon_j)|X_iX_j)=0$，$i\neq j$。

假定 5：解释变量之间没有完全的多重共线性(仅适用于多元模型)。

说明：违反第一条假定，即 X 是随机变量且同时和随机误差项相关，即出现了随机解释变量问题，就会违反第三条假定，即出现了异方差性，也就会违反第四条和第五条假定，即出现了序列相关性和多重共线性。

另外，在模型估计时，在前述 5 条假定的基础上还可以加上以下的一些假定：

假定 6：回归模型对参数是线性的。

假定 7：样本容量必须大于待估计的参数个数。

假定 8：模型设定是正确的。

假定 9：$\mathrm{cov}(\varepsilon_t,x_t)=0$，即误差与变量是独立的。

假定 10：随机误差项服从正态分布。

5.3.2 多元线性回归模型的矩阵解法

当已知 n 组自变量 $x_j(j=1,2,\cdots,m)$和因变量 y 的观测值时，则可写出 n 个方程式的方程组，其中未知数为 $m+1$ 个回归系数。该方程组可写成矩阵形式：

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

$$\boldsymbol{Y}=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},\quad \boldsymbol{X}=\begin{bmatrix}1&x_{11}&\cdots&x_{m1}\\1&x_{12}&\cdots&x_{m2}\\\vdots&\vdots&\ddots&\vdots\\1&x_{1n}&\cdots&x_{mn}\end{bmatrix},\quad \boldsymbol{\beta}=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_m\end{bmatrix},\quad \boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}$$

因为 $\boldsymbol{X}$ 矩阵中一般 $n\neq m$，故 $\boldsymbol{X}$ 无法求逆；为求解 $\boldsymbol{\beta}$ 可两边左乘 $\boldsymbol{X}^{\mathrm{T}}$，得

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}$$

而 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 为方阵，可求逆，则可得

$$\boldsymbol{\beta}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_m\end{bmatrix}$$

5.3.3 多元线性回归模型的统计检验

1. 标准离差检验

(1) 因变量标准差 S 检验。

$$S=\sqrt{\frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}}{n-m-1}}\tag{5-8}$$

(2) 各个回归系数标准差 $S_{\beta i}(i=0,1,2,\cdots,m)$的检验。

$$S_{\beta i}=\sqrt{C_{ii}}S(i=0,1,\cdots,m)\tag{5-9}$$

式中 C_{ii}——$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$矩阵主对角线上的第 i 项的值。

2. 相关系数检验

$$R=\sqrt{\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}-n\bar{y}^{2}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-n\bar{y}^{2}}} \tag{5-10}$$

R 值愈接近 1 愈好，可查相关系数检验表。

为了说明多元线性回归线对样本观测值的拟合情况，可以考察在 $\boldsymbol{Y}$ 的总变差中有多个解释变量做出了解释的那部分变差的比重，即回归平方和与总离差平方和的比重，在多元回归中这一比重称为多重可决系数，用 R^2 表示。

1）变差

$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y})^2$$

模型所要解释的是 y 相对于其均值的波动性。

总离差平方和 = 回归平方和 + 残差平方和

$$\mathrm{TSS}=\mathrm{RSS}+\mathrm{ESS}$$

2）自由度

$$n-1=(n-k)+(k-1)$$

定义

$$R^2=\frac{\mathrm{ESS}}{\mathrm{TSS}}$$

$$R^2=\frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}}=1-\frac{\mathrm{RSS}}{\mathrm{TSS}}=1-\frac{\sum e_i^2}{\sum(y_i-\bar{y})^2}$$

多重可决系数可用矩阵表示，因为

$$\mathrm{TSS}=Y'Y-n\bar{y}^2$$

$$\mathrm{ESS}=\hat{\beta}X'Y-n\bar{y}^2$$

所以

$$R^2=\frac{\mathrm{ESS}}{\mathrm{TSS}}=\frac{\hat{\beta}X'Y-n\bar{y}^2}{Y'Y-n\bar{y}^2}$$

由此可知：

$$R^2=\frac{\hat{\beta}X'Y-n\bar{y}^2}{Y'Y-n\bar{y}^2}=\frac{\hat{\beta}_2\sum x_{2i}y_i+\hat{\beta}_3\sum x_{3i}y_i+\cdots+\hat{\beta}_k\sum x_{ki}y_i}{\sum y_i^2}$$

可见多重可决系数是模型中解释变量个数的不减函数，也就是说，随着模型中解释变量的增加，多重可决系数 R^2 的值会增大。当被解释变量相同而解释变量个数不同时，会给运用多重可决系数比较两个模型的拟合程度带来缺陷。这时模型的解释变量个数不同，不能简单地直接对比多重可决系数。可决系数只涉及变差，没有考虑自由度（可自由变化的样本观测个数，等于所用样本观测值的个数减去对观测值的约束个数）。显然，如果用自由度去校正所计算的变差，可以纠正解释变量个数不同引起的对比困难。因为在样本容量一定的情况下，增加解释变量必定使得待估参数的个数增加，从而会损失自由度，为此可以用自由度去修正多重可决系数中的残差平方和与回归平方和，有：

$$R^2=1-\frac{\sum e_i^2/(n-k)}{\sum(y_i-\bar{y})^2/(n-1)}=1-\frac{n-1}{n-k}\frac{\sum e_i^2}{\sum(y_i-\bar{y})^2}$$

修正可决系数与可决系数的关系如下：

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

可见：$k>1$ 时，$\bar{R}^2<R^2$，这意味着随着解释变量的增加，$\bar{R}^2<R^2$，若 $\bar{R}^2$ 为负数时，规定 $\bar{R}^2=0$。

3. *F* 检验

F 检验用来检验多元线性回归模型的总体效果。

$$F=\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}Y}{mS^2}=\frac{\mathrm{ESS}/m}{\mathrm{RSS}/(n-m-1)}\sim F(m,n-m-1) \tag{5-11}$$

计算出 F 值后，再查 F 检验表得 F_α，当 $F\geqslant F_\alpha$ 时，检验通过，模型有效。

4. *t* 检验

t 检验用来检验回归系数 $\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_m$ 的统计意义，即检验自变量 $x_1,x_2,\cdots,x_m$ 对 y 的影响显著与否。

$$t_{\beta i}=\hat{\beta}_i/S_{\beta i}(i=1,2,\cdots,m)\sim t(n-m-1)$$

按此式算出 $t_{\beta i}$ 值后，再查 t 检验表得 $t_{\alpha/2}$。$t_{\beta i}>t_{\alpha/2}$ 时，检验通过；否则，应剔除相应的自变量。

5. 预测置信区间的确定

按正态分布理论，当取置信度为 95%时，预测值置信区间为

$$\hat{y}=\bar{y}_0\pm 2S$$

要注意的是，在多元线性回归模型的构建中，可能会遇到多重共线性的问题。多重共线性是指自变量之间存在线性关系或接近线性关系。如果它们完全相关，则$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$不存在，最小二乘法就失效了；应用最小二乘法估计回归系数的一个重要条件就是自变量之间为不完全的线性相关。如果这种相关程度较低，其影响可以忽略；但若高度相关，则回归系数无效或无意义，因而所建模型无效。这时应选择其他新的自变量以替换相关的变量或采用其他方法来建立模型。关于这部分内容，将在后面的章节讨论。

5.4 多元线性回归分析的 R 语言应用

例 5-3：为了检验美国电力行业是否存在规模经济，Nerlove(1963)搜集了 1955 年 145 家美国电力企业的总成本(TC)、产量(Q)、工资率(PL)、燃料价格(PF)及资本租赁价格(PK)的数据，如表 5-4 所示。试以总成本为因变量，以产量、工资率、燃料价格和资本租赁价格为自变量，利用多元线性回归分析方法研究它们之间的关系。

表 5-4 美国电力行业数据

编号	TC/百万美元	Q/千瓦时	PL/美元/千瓦时	PF/美元/千瓦时	PK/美元/千瓦时
1	0.082	2	2.09	17.9	183
2	0.661	3	2.05	35.1	174
3	0.99	4	2.05	35.1	171
4	0.315	4	1.83	32.2	166

续表

编号	TC/百万美元	Q/千瓦时	PL/美元/千瓦时	PF/美元/千瓦时	PK/美元/千瓦时
5	0.197	5	2.12	28.6	233
6	0.098	9	2.12	28.6	195
⋮	⋮	⋮	⋮	⋮	⋮
143	73.05	11796	2.12	28.6	148
144	139.422	14359	2.31	33.5	212
145	119.939	16719	2.3	23.6	162

在目录G:\2glkx\data下建立al5-2.xls数据文件后,使用的命令如下:

```
> library(RODBC)  #使用此命令时必须先安装RODBC,见3.9.2节
> z<-odbcConnectExcel("G:/2glkx/data/al5-2.xls")
> sq<-sqlFetch(z,"Sheet1")
> close(z)
> sq
```

执行以上5行命令后,得到如下结果:

```
        TC       Q      PL     PF     PK
 1    0.082      2    2.09   17.9   183
 ⋮
145  119.939  16719  2.30   23.6   162
```

1. 对数据进行描述性分析

在符号“>”后输入如下命令:

```
> y<-sq$TC; x1<-sq$Q; x2<-sq$PL; x3<-sq$PF; x4<-sq$PK
> d<-data.frame(y,x1,x2,x3,x4)
> summary(d)
#这些命令的含义是对总成本(TC)、产量(Q)、工资率(PL)、燃料价格(PF)及资本租赁价格(PK)等变量进行详细的描述性分析
```

输入以上3行命令后,按回车键,得到如下分析结果:

```
       y                  x1               x2              x3
Min.   :   0.082   Min.   :     2   Min.   : 1.450   Min.   :10.30
1st Qu.:   2.382   1st Qu.:   279   1st Qu.: 1.760   1st Qu.:21.30
Median :   6.754   Median :  1109   Median : 2.040   Median :26.90
Mean   :  12.976   Mean   :  2133   Mean   : 1.972   Mean   :26.18
3rd Qu.:  14.132   3rd Qu.:  2507   3rd Qu.: 2.190   3rd Qu.:32.20
Max.   : 139.422   Max.   : 16719   Max.   : 2.320   Max.   :42.80
       x4
Min.   :138.0
1st Qu.:162.0
Median :170.0
Mean   :174.5
3rd Qu.:183.0
Max.   :233.0
```

通过观察上面的结果，可以得到很多信息，如5个最小值、第一百分位数、中位数、平均值、最大值等。更多的信息描述如下。

(1) 5个最小值(Smallest)。

总成本(TC)最小值是0.082。

产量(Q)最小值是2。

工资率(PL)最小值是1.450。

燃料价格(PF)最小值是10.30。

资本租赁价格(PK)最小值是138.0。

(2) 百分位数。

5个变量的第一百分位数分别是2.382,279,1.760,21.30,162.0。

5个变量的第三百分位数分别是14.132,2507,2.190,32.20,233.0。

(3) 5个中位数(median)。

5个变量的中位数分别是6.754,1109,2.040,26.90,170.0。

(4) 5个平均值(Mean)。

5个变量的平均值分别是12.976,2133,1.972,26.18,174.5。

(5) 5个最大值(Largest)。

5个变量的最大值分别是139.422,16719,2.320,42.80,233.0。

2. 对数据进行相关分析

在符号“>”后输入如下命令：

```
> d <- data.frame(y,x1,x2,x3,x4)
> cor(d)
```

输入完以上2行命令后，按回车键，得到如下分析结果。

```
            y           x1          x2          x3          x4
y   1.00000000  0.952503699  0.2513375  0.03393519  0.027202000
x1  0.95250370  1.000000000  0.1714499 -0.07734943  0.002869139
x2  0.25133754  0.171449901  1.0000000  0.31370293 -0.178145470
x3  0.03393519 -0.077349434  0.3137029  1.00000000  0.125428217
x4  0.02720200  0.002869139 -0.1781455  0.12542822  1.000000000
```

通过观察上面的结果，可以看到变量TC与各个变量之间的相关关系还是可以接受的，可以进行下面的回归分析。

3. 对数据进行回归分析

在符号“>”后输入如下命令：

```
> lm.reg <- lm(y~1 + x1 + x2 + x3 + x4)
# 本命令是对总成本(TC)、产量(Q)、工资率(PL)、燃料价格(PF)及资本租赁价格(PK)等变量进行多元回归分析
> summary(lm.reg)
```

每输入完一条命令后，按回车键，最后得到如下分析结果：

```
Call:
lm(formula = y ~ 1 + x1 + x2 + x3 + x4)
Residuals:
   Min       1Q   Median     3Q    Max
-17.814  -1.609  -0.092  2.231  43.761
    Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.222e+01  6.587e+00  -3.373  0.000961  ***
x1            6.395e-03  1.629e-04  39.258  < 2e-16  ***
x2            5.655e+00  2.176e+00   2.598  0.010366  *
x3            2.078e-01  6.410e-02   3.242  0.001482  **
x4            2.844e-02  2.650e-02   1.073  0.285088
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.579 on 140 degrees of freedom
Multiple R-squared: 0.9228, Adjusted R-squared: 0.9206
F-statistic: 418.1 on 4 and 140 DF, p-value: < 2.2e-16
```

通过观察上面的分析结果，可以看出：模型的 F 值＝418.12，p 值＝0.0000，说明模型整体上是非常显著的。模型的可决系数 R-squared＝0.9228，修正的可决系数 Adjusted R-squared ＝0.9206，说明模型的解释能力是可以的。

模型的回归方程是

$$TC=0.006395Q+5.655PL+0.2078PF+0.02844PK-22.22$$

变量 Q 系数标准误是 0.0001629，t 值为 39.258，p 值为 0.000，系数是非常显著的。变量 PL 系数标准误是 2.176，t 值为 2.598，p 值为 0.010，系数是非常显著的。变量 PF 系数标准误是 0.06410，t 值为 3.242，p 值为 0.001，系数是非常显著的。变量 PK 系数标准误是 0.0265，t 值为 1.073，p 值为 0.285，系数是非常不显著的。常数项的系数标准误是 22.22，t 值为－3.373，p 值为 0.000961，系数是非常显著的。

综合上面的分析，可以看出：美国电力企业的总成本(TC)受到产量(Q)、工资率(PL)、燃料价格(PF)、资本租赁价格(PK)的影响，美国电力行业存在规模经济特征。

应注意上面的模型中 PK 的系数是不显著的，下面把该变量剔除后重新进行回归分析，命令如下：

```
> lm.reg<-lm(y~1+x1+x2+x3)
> summary(lm.reg)
```

输入完以上 2 行命令后，按回车键，则得到如下分析结果：

```
Call:
lm(formula = y ~ 1 + x1 + x2 + x3)
Residuals:
     Min       1Q   Median     3Q    Max
-17.290  -1.503  -0.385  2.179  44.779
Coefficients:
             Estimate Std.        Error  t value  Pr(>|t|)
(Intercept)  -1.654e+01  3.928e+00  -4.212  4.48e-05  ***
x1            6.406e-03  1.627e-04  39.384  < 2e-16  ***
x2            5.098e+00  2.115e+00   2.411  0.017208   *
x3            2.217e-01  6.283e-02   3.528  0.000565  ***
```

```
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.582 on 141 degrees of freedom
Multiple R-squared: 0.9221, Adjusted R-squared: 0.9205
F-statistic: 556.5 on 3 and 141 DF, p-value: < 2.2e-16
>q()     #退出R
```

从上面分析结果可见，模型整体依旧是非常显著的。模型的可决系数以及修正的可决系数变化不大，说明模型的解释能力几乎没有变化。其他变量（含常数项的系数）都非常显著，模型接近完美。可以把回归结果作为最终的回归模型方程，即

$$TC=0.006406Q+5.098PL+0.2217PF-16.54$$

从上面的分析可以看出，美国电力企业的总成本受到产量、工资率、燃料价格的影响。总成本随着这些变量值的升高而升高、降低而降低。

值得注意的是：产量的增加引起总成本的相对变化是很小的，所以，从经济意义上说，美国电力行业存在规模经济特征。

5.5 稳健线性回归分析的R语言应用

5.5.1 线性回归中的几个术语

先介绍几个线性回归（linear regression）中的术语。

（1）残差（residual）：基于回归方程的预测值与观测值的差。

（2）离群点（outlier）：线性回归中的离群点是指对应残差较大的观测值。也就是说，当某个观测值与基于回归方程的预测值相差较大时，该观测值即可视为离群点。离群点的出现一般是因为样本自身较为特殊或者数据录入错误导致的，当然也可能是其他问题。

（3）杠杆率（leverage）：当某个观测值所对应的预测值为极端值时，该观测值称为高杠杆率点。杠杆率衡量的是独立变量对自身均值的偏异程度。高杠杆率的观测值对于回归方程的参数有重大影响。

（4）影响力点（influence）：若某观测值的剔除与否对回归方程的系数估计有显著影响，则该观测值是具有影响力的，称为影响力点。影响力是高杠杆率和离群情况引起的。

（5）Cook距离（Cook's distance）：综合了杠杆率信息和残差信息的统计量。

使用最小二乘回归时，有时候会遇到离群点和高杠杆率点。此时，若认定离群点或者高杠杆率点的出现并非因为数据录入错误或者该观测值来自另外一个总体的话，使用最小二乘回归会变得很棘手，因为数据分析者因为没有充分的理由剔除离群点和高杠杆率。此时稳健回归是极佳的替代方案。稳健回归在剔除离群点或者高杠杆率点和保留离群点或高杠杆率点并像最小二乘法那样平等使用各点之间找到了一个折中。其在估计回归参数时，根据观测值的稳健情况对观测值进行赋权。简言之，稳健回归是加权最小二乘回归，或称稳健最小二乘回归。

MASS包中的rlm命令提供了不同形式的稳健回归拟合方式。接下来，以基于Huber方法和bisquare方法下的M估计为例来进行演示。这是两种最为基本的M估计方法。在M估计中，要做的事情是在满足约束$\sum_{i=1}^{n}w_i(y_i-x'b)x'_i=0$时，求出使得$\sum_{i=1}^{n}w_i^2e_i^2$最小的参

数。由于权重的估计依赖于残差，而残差的估计又反过来依赖于权重。因此，需用迭代重复加权最小二乘(Iteratively Reweighted Least Squares，IRLS)来估计参数。举例来说，第 j 次迭代得到的系数矩阵为 $\boldsymbol{B}_j=[\boldsymbol{X}'w_{j-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'w_{j-1}\boldsymbol{Y}$，这里下标表示求解过程中的迭代次数，而不是通常的行标或者列标，持续这一过程，直到结果收敛为止。在 Huber 方法下，残差较小的观测值被赋予的权重为 1，残差较大的观测值的权重随着残差的增大而递减。而在 bisquare 方法下，所有的非 0 残差所对应的观测值的权重都是递减的。

5.5.2 数据描述

下面用到的数据是 Alan Agresti 和 Barbara Finlay 所著的 *Statistical Methods for Social Sciences*(Third Edition，Prentice Hall，1997)中的 crime 数据集。该数据集的变量分别是

```
state id (sid),
state name (state),
violent crimes per 100,000 people (crime),
murders per 1,000,000 (murder),
the percent of the population living in metropolitan areas (pctmetro),
the percent of the population that is white (pctwhite),
percent of population with a high school education or above (pcths),
percent of population living under poverty line (poverty),
percent of population that are single parents (single).
```

该数据集共有 51 个观测值。接下来用数据集中的 poverty 和 single 变量来预测 crime：

```
> ibrary(foreign)
> cdata <- read.dta("http://www.ats.ucla.edu/stat/data/crime.dta")
> summary(cdata)
```

执行以上 3 行命令后，得到如下结果：

```
      sid             state              crime          murder
Min.   :  1.0   Length:         51   Min.   :   82.0   Min.   : 1.600
1st Qu.: 13.5   Class :  character   1st Qu.:  326.5   1st Qu.: 3.900
Median : 26.0   Mode  :  character   Median :  515.0   Median : 6.800
Mean   : 26.0                        Mean   :  612.8   Mean   : 8.727
3rd Qu.: 38.5                        3rd Qu.:  773.0   3rd Qu.:10.350
Max.   : 51.0                        Max.   : 2922.0   Max.   :78.500
    pctmetro        pctwhite         pcths          poverty
Min.   : 24.00   Min.   : 31.80   Min.   : 64.30   Min.   : 8.00
1st Qu.: 49.55   1st Qu.: 79.35   1st Qu.: 73.50   1st Qu.:10.70
Median : 69.80   Median : 87.60   Median : 76.70   Median :13.10
Mean   : 67.39   Mean   : 84.12   Mean   : 76.22   Mean   :14.26
3rd Qu.: 83.95   3rd Qu.: 92.60   3rd Qu.: 80.10   3rd Qu.:17.40
Max.   :100.00   Max.   : 98.50   Max.   : 86.60   Max.   :26.40
     single
Min.   : 8.40
1st Qu.:10.05
```

```
Median :10.90
Mean   :11.33
3rd Qu.:12.05
Max.   :22.10
```

5.5.3 普通最小二乘(OLS)回归的R语言应用

先对数据进行普通最小二乘(OLS)回归,重点观察回归结果中的残差、拟合值、Cook距离和杠杆率。

```
> ols <- lm(crime~poverty + single, data = cdata)
> summary(ols)
```

执行以上2行命令后,得到如下结果:

```
Call:
lm(formula = crime ~ poverty + single, data = cdata)
Residuals:
   Min        1Q      Median     3Q      Max
-811.14   -114.27   -22.44   121.86   689.82
Coefficients:
               Estimate  Std. Error   t value    Pr(>|t|)
(Intercept)   -1368.189     187.205   -7.308   2.48e-09   ***
poverty           6.787       8.989    0.755       0.454
single          166.373      19.423    8.566   3.12e-11   ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 243.6 on 48 degrees of freedom
Multiple R-squared: 0.7072, Adjusted R-squared: 0.695
F-statistic: 57.96 on 2 and 48 DF, p-value: 1.578e-13
```

执行以下命令:

```
> opar <- par(mfrow = c(2,2),oma = c(0, 0, 1.1, 0))
> plot(ols, las = 1)
```

得到如图5-1所示的结果。

从图5-1中可以看出,第9、第25和第51个观测值可能是离群点,看看这些观测值所属的是美国的哪些州。

```
> cdata[c(9, 25, 51), 1:2]
```

执行以上命令后,得到如下结果:

```
   sid state
9    9    fl
25  25    ms
51  51    dc
```

可以猜测,DC、Florida和Mississippi这三个州所对应的观测值可能具有较大的残差或者杠杆率。

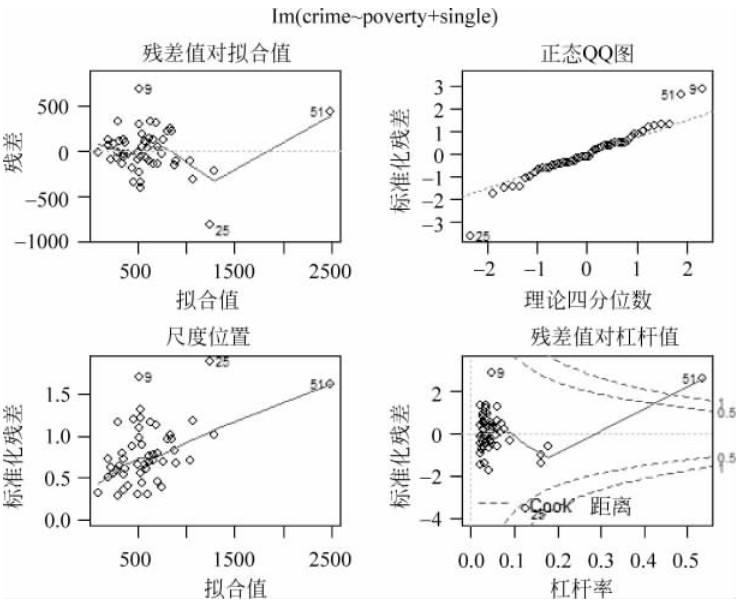

图 5-1　残差、拟合值、Cook 距离和杠杆率

下面观察 Cook 距离较大的观测值有哪些。在判断 Cook 距离大小的时候，通常采用的经验分界点是 Cook 距离序列的 $4/n$ 处，其中 n 是观测值的个数。

```
> library(MASS)
> d1 <- cooks.distance(ols)
> r <- stdres(ols)
> a <- cbind(cdata, d1, r) a[d1 > 4/51, ]
```

执行以上 4 行命令后，得到如下结果：

```
错误: 意外的符号 in "a <- cbind(cdata, d1, r) a"
> a <- cbind(cdata,d1,r)
> a[d1 > 4/51,]
```

执行以上 2 行命令后，得到如下结果：

```
    sid  state  crime  murder  pctmetro  pctwhite  pcths  poverty  single         d1
1     1     ak    761     9.0      41.8      75.2   86.6      9.1    14.3  0.1254750
9     9     fl   1206     8.9      93.0      83.5   74.4     17.8    10.6  0.1425891
25   25     ms    434    13.5      30.7      63.3   64.3     24.7    14.7  0.6138721
51   51     dc   2922    78.5     100.0      31.8   73.1     26.4    22.1  2.6362519
             r
1   -1.397418
9    2.902663
25  -3.562990
51   2.616447
```

本来应当先删除 DC 所对应的观测值，因为 DC 对应的并不是州。然而，由于 DC 所对应的 Cook 距离较大，所以保留 DC 有助于观察。

下面生成一个 absr1 变量，其对应的为残差序列的绝对值，取出残差绝对值较大的观测值：

```
> rabs <- abs(r)
> a <- cbind(cdata, d1, r, rabs)
> asorted <- a[order(-rabs), ]
> asorted[1:10, ]
```

执行以上4行命令后,得到如下结果:

```
    sid  state  crime  murder  pctmetro  pctwhite  pcths  poverty  single         d1
25   25     ms    434    13.5      30.7      63.3   64.3     24.7    14.7  0.61387212
9     9     fl   1206     8.9      93.0      83.5   74.4     17.8    10.6  0.14258909
51   51     dc   2922    78.5     100.0      31.8   73.1     26.4    22.1  2.63625193
46   46     vt    114     3.6      27.0      98.4   80.8     10.0    11.0  0.04271548
26   26     mt    178     3.0      24.0      92.6   81.0     14.9    10.8  0.01675501
21   21     me    126     1.6      35.7      98.5   78.8     10.7    10.6  0.02233128
1     1     ak    761     9.0      41.8      75.2   86.6      9.1    14.3  0.12547500
31   31     nj    627     5.3     100.0      80.8   76.7     10.9     9.6  0.02229184
14   14     il    960    11.4      84.0      81.0   76.2     13.6    11.5  0.01265689
20   20     md    998    12.7      92.8      68.9   78.4      9.7    12.0  0.03569623
             r      rabs
25  -3.562990  3.562990
9    2.902663  2.902663
51   2.616447  2.616447
46  -1.742409  1.742409
26  -1.460885  1.460885
21  -1.426741  1.426741
1   -1.397418  1.397418
31   1.354149  1.354149
14   1.338192  1.338192
20   1.287087  1.287087
```

5.5.4 稳健回归的R语言应用

现在转向稳健回归。

再解释一下,稳健回归是通过迭代重复加权最小二乘(IRLS)来完成的。其对应的R语言函数是MASS包中的rlm()。IRLS对应的有多个权重函数(weighting functions),首先演示一下Huber方法。演示过程中,重点关注IRLS过程得出的权重结果。

```
> rr.huber <- rlm(crime ~ poverty + single, data = cdata)
> summary(rr.huber)
```

执行以上2行命令,得到如下结果:

```
Call: rlm(formula = crime ~ poverty + single, data = cdata)
Residuals:
     Min       1Q   Median       3Q      Max
-846.09  -125.80   -16.49   119.15   679.94
Coefficients:
                  Value   Std. Error   t value
(Intercept)   -1423.0373    167.5899    -8.4912
poverty           8.8677      8.0467     1.1020
single          168.9858     17.3878     9.7186
```

```
Residual standard error: 181.8 on 48 degrees of freedom

> hweights <- data.frame(state = cdata$state, resid = rr.huber$resid, weight = rr.huber
$w)
> hweights2 <- hweights[order(rr.huber$w), ]
> hweights2[1:15, ]
```

执行“>”符号后的3行命令后，得到如下结果：

```
    state        resid     weight
25     ms  -846.08536  0.2889618
9      fl   679.94327  0.3595480
46     vt  -410.48310  0.5955740
51     dc   376.34468  0.6494131
26     mt  -356.13760  0.6864625
21     me  -337.09622  0.7252263
31     nj   331.11603  0.7383578
14     il   319.10036  0.7661169
1      ak  -313.15532  0.7807432
20     md   307.19142  0.7958154
19     ma   291.20817  0.8395172
18     la  -266.95752  0.9159411
2      al   105.40319  1.0000000
3      ar    30.53589  1.0000000
4      az   -43.25299  1.0000000
```

容易看出，观测值的残差绝对值越大，其被赋予的权重越小。结果表明：Mississippi所对应的观测值被赋予的权重是最小的，其次是Florida所对应的观测值，而所有未被展示的观测值的权重均为1。由于OLS回归中所有观测值的权重都为1，因此，稳健回归中权重为1的观测值越多，则稳健回归与OLS回归的分析结果越相近。

接下来，用bisquare方法来实现稳健回归过程。

```
> rr.bisquare <- rlm(crime ~ poverty + single, data = cdata, psi = psi.bisquare)
> summary(rr.bisquare)
```

执行以上2行命令后，得到如下结果：

```
Call: rlm(formula = crime ~ poverty + single, data = cdata, psi = psi.bisquare)
Residuals:
     Min       1Q   Median       3Q      Max
-905.59  -140.97   -14.98   114.65   668.38
Coefficients:
                  Value   Std. Error   t value
(Intercept)   -1535.3338    164.5062    -9.3330
poverty          11.6903      7.8987     1.4800
single          175.9303     17.0678    10.3077
Residual standard error: 202.3 on 48 degrees of freedom

> biweights <- data.frame(state = cdata$state, resid = rr.bisquare$resid, weight = rr.
```

```
bisquare$w)
> biweights2 <- biweights[order(rr.bisquare$w), ]
> biweights2[1:15, ]
```

执行以上3行命令后，得到如下结果：

```
    state       resid       weight
25     ms  -905.5931  0.007652565
9      fl   668.3844  0.252870542
46     vt  -402.8031  0.671495418
26     mt  -360.8997  0.731136908
31     nj   345.9780  0.751347695
18     la  -332.6527  0.768938330
21     me  -328.6143  0.774103322
1      ak  -325.8519  0.777662383
14     il   313.1466  0.793658594
20     md   308.7737  0.799065530
19     ma   297.6068  0.812596833
51     dc   260.6489  0.854441716
50     wy  -234.1952  0.881660897
5      ca   201.4407  0.911713981
10     ga  -186.5799  0.924033113
```

与Huber方法相比，bisquare方法下的Mississippi观测值被赋予了极小的权重，并且两种方法估计出的回归参数也相差甚大。通常，当稳健回归与OLS回归的分析结果相差较大时，数据分析者采用稳健回归较为明智。稳健回归和OLS回归的分析结果的较大差异通常暗示着离群点对模型参数产生了较大影响。所有的方法都有长处和短处，稳健回归也不例外。在稳健回归中，Huber方法的短处在于无法很好地处理极端离群点，而bisquare方法的短处在于回归结果不易收敛，以至于经常有多个最优解。

除此之外，两种方法得出的参数结果极为不同，尤其是single变量的系数和截距项(intercept)。不过，一般而言无须关注截距项，除非事先已经对预测变量进行了中心化，此时截距项才显得有些用处。再有，变量poverty的系数在两种方法下都不显著，而变量single则刚好相反，都较为显著。

练习题

1. 为了给今后编制管理费用的预算提供数据，某企业分析了近10年来企业管理费用与产值之间的关系，如表5-5所示。

表5-5 数据表

年份	1	2	3	4	5	6	7	8	9	10
管理费用/百万元	5.9	6.3	6.5	7.3	6.9	7.8	8.5	8.1	9.2	9.4
产值/千万元	5.2	5.8	6.3	6.8	7.5	8.3	9.1	10.0	10.9	11.8

(1) 使用R语言建立该企业管理费用与产值之间的线性回归模型，求出回归方程并进行检验；(2)下一年该企业的产值预计为1.5亿元，使用R语言求管理费用的置信度为95%的预测区间。

2. 某电子集团公司分析企业的劳动生产率和企业在研究与开发(R&D)投入之间的关系，调查了下属14个企业2002年的劳动生产率与R&D投入占销售额的比例数据，如表5-6所示。

表5-6　数据表

R&D投入占销售额比例/%	1.4	1.4	1.5	1.4	1.7	2.0	2.0
劳动生产率/万元/人	6.7	6.9	7.2	7.3	8.4	8.8	9.1
R&D投入占销售额比例/%	2.4	2.5	2.6	2.7	2.8	3.1	3.5
劳动生产率/万元/人	9.8	10.6	10.7	11.1	11.8	12.1	13.0

(1) 劳动生产率与R&D投入比例之间是否呈线性相关关系(使用R语言散点图分析)？若是，使用R语言求它们之间的回归方程；

(2) 该集团企业的R&D投入率为4.6%，使用R语言求该企业劳动生产率的置信度为90%的预测区间。

3. 为了研究深圳市地方预算内财政收入与国内生产总值的关系，得到如表5-7所示的数据。

表5-7　深圳市地方预算内财政收入与国内生产总值

年份	地方预算内财政收入Y/亿元	国内生产总值(GDP)X/亿元
1990	21.7037	171.6665
1991	27.3291	236.6630
1992	42.9599	317.3194
1993	67.2507	449.2889
1994	74.3992	615.1933
1995	88.0174	795.6950
1996	131.7490	950.0446
1997	144.7709	1130.0133
1998	164.9067	1289.0190
1999	184.7908	1436.0267
2000	225.0212	1665.4652
2001	265.6532	1954.6539

资料来源：《深圳统计年鉴2002》，中国统计出版社。

(1) 使用R语言建立深圳地方预算内财政收入对GDP的回归模型；

(2) 估计所建立模型的参数，解释斜率系数的经济意义；

(3) 对回归结果进行检验；

(4) 若2005年的国内生产总值为3600亿元,确定2005年财政收入的预测值和预测区间($\alpha=0.05$)。

4. 某企业研究与发展经费与利润的数据如表5-8所示。

表5-8 某企业研究与发展经费与利润的数据

年份	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
研究与发展经费/万元	10	10	8	8	8	12	12	12	11	11
利润额/万元	100	150	200	180	250	300	280	310	320	300

使用R语言分析企业研究与发展经费与利润额的相关关系,并作回归分析。

5. 为研究中国的货币供应量(以货币与准货币M2表示)与国内生产总值的相互依存关系,分析表5-9中1990—2001年中国货币供应量和国内生产总值的有关数据。

表5-9 货币供应量与国内生产总值数据

年份	货币供应量/亿元	国内生产总值/亿元
1990	1529.3	18598.4
1991	19349.9	21662.5
1992	25402.2	26651.9
1993	34879.8	34560.5
1994	46923.5	46670.0
1995	60750.5	57494.9
1996	76094.9	66850.5
1997	90995.3	73142.7
1998	104498.5	76967.2
1999	119897.9	80579.4
2000	134610.3	88228.1
2001	158301.9	94346.4

资料来源:《中国统计年鉴2002》,第51页、第662页,中国统计出版社。

使用R语言对货币供应量与国内生产总值作相关分析,并说明分析结果的经济意义。

6. 表5-10是16支公益股票某年的每股账面价值和当年红利。

表5-10 16支公益股票某年的每股账面价值和当年红利

公司序号	账面价值/元	红利/元	公司序号	账面价值/元	红利/元
1	22.44	2.4	9	12.14	0.80
2	20.89	2.98	10	23.31	1.94
3	22.09	2.06	11	16.23	3.00
4	14.48	1.09	12	0.56	0.28
5	20.73	1.96	13	0.84	0.84
6	19.25	1.55	14	18.05	1.80
7	20.37	2.16	15	12.45	1.21
8	26.43	1.60	16	11.33	1.07

根据表5-10的资料：

(1) 使用R语言建立每股账面价值和当年红利的回归方程；

(2) 解释回归系数的经济意义；

(3) 若序号为6的公司的股票每股账面价值增加1元，估计当年红利可能为多少。

7. 从某工业部门抽取10个生产单位进行调查，得到如表5-11所示的数据。

表5-11　某工业部门年产量和工作人数

单位序号	年产量/万吨	工作人员数/千人
1	210.8	7.062
2	210.1	7.031
3	211.5	7.018
4	208.9	6.991
5	207.4	6.974
6	205.3	7.953
7	198.8	6.927
8	192.1	6.302
9	183.2	6.021
10	176.8	5.310

要求：假定年产量与工作人员数之间存在线性关系，试用经典回归估计该工业部门的生产函数及边际劳动生产率。

8. 表5-12给出了1988年9个工业国的名义利率(Y)与通货膨胀率(X)的数据。

表5-12　名义利率(Y)与通货膨胀率(X)的数据

国家	Y/%	X/%
澳大利亚	11.9	7.7
加拿大	9.4	4.0
法国	7.5	3.1
德国	4.0	1.6
意大利	11.3	4.8
墨西哥	66.3	51.0
瑞典	2.2	2.0
英国	10.3	6.8
美国	7.6	4.4

资料来源：原始数据来自国际货币基金组织出版的《国际金融统计》。

要求：(1) 使用R语言，以利率为纵轴、通货膨胀率为横轴做图；

(2) 使用R语言进行回归分析；

(3) 如果实际利率不变，则名义利率与通货膨胀率的关系如何？

9. 现代投资分析的特征线涉及如下回归方程：$r_t=\beta_0+\beta_1 r_{mt}+u_t$；其中，$r$表示股票或债券的收益率；$r_m$表示有价证券的收益率(用市场指数表示，如标准普尔500指数)；t表示时间。在投资分析中，β_1被称为债券的安全系数β，是用来度量市场的风险程度的，即市场的发展对公司的财产有何影响。依据1956—1976年间240个月的数据，Fogler和Ganpathy

得到IBM股票的回归方程；市场指数是在芝加哥大学建立的市场有价证券指数：

$$\hat{r}_t = 0.7264 + 1.0598 r_{mt} \qquad r^2 = 0.4710$$
$$(0.3001)\ (0.0728)$$

要求：(1) 解释回归参数的意义；

(2) 如何解释 r^2？

(3) 安全系数 $\beta>1$ 的证券称为不稳定证券，建立适当的零假设及备选假设，并用 t 检验进行检验($\alpha=5\%$)。

10. 在某种钢材的试验中，研究了延伸率 Y(%)与含碳量 X_1(单位0.01%)及回火温度 X_2 之间的关系，表5-13给出了15批生产试验数据。

(1) 求延伸率与含碳量、回火温度之间的二元线性回归方程，并分析软件运行输出结果；

(2) 要求以90%的把握将该钢材的延伸率控制在15%以上，问当含碳量为60(单位0.01%)时，应将回火温度控制在哪一范围内？

表5-13 数 据

Y_i/%	19.25	17.50	18.25	16.25	17.00	16.75	17.00	16.75
X_{i1}/0.01%	57	64	69	58	58	58	58	58
X_{i2}/℃	535	535	535	460	460	460	490	490
Y_i/%	17.25	16.75	14.75	12.00	17.75	17.75	15.50	
X_{i1}/0.01%	58	57	64	69	59	64	69	
X_{i2}/℃	490	460	435	460	490	467	490	

11. 一般认为，一个地区的农业总产值与该地区的农业劳动力、灌溉面积、施用化肥量、农户固定资产以及农业机械化水平诸因素有很大关系。表5-14给出了1985年我国北方地区12个省市的农业总产值与农业劳动力、灌溉面积、化肥用量、户均固定资产、农机动力的调查数据。

表5-14 我国北方地区农业投入和产出数据

地区	农业总产值/亿元	农业劳动力/万人	灌溉面积/万公顷	化肥用量/万吨	户均固定资产/元	农机动力/万马力
北京	19.61	90.1	33.84	7.5	394.30	435.3
天津	14.40	95.2	34.95	3.9	567.50	450.7
河北	149.90	1639.0	357.26	92.4	706.89	2712.6
山西	55.07	562.6	107.90	31.4	856.37	1118.5
内蒙古	60.85	462.9	96.49	15.4	1282.81	641.7
辽宁	87.48	588.9	72.40	61.6	844.74	1129.6
吉林	73.81	399.7	69.63	36.9	2576.81	647.6
黑龙江	104.51	425.3	67.95	25.8	1237.16	1305.8
山东	276.55	2365.6	456.55	152.3	5812.02	3127.9
河南	200.02	2557.5	318.99	127.9	754.78	2134.5
陕西	68.18	884.2	117.90	36.1	607.41	764.0
新疆	49.12	256.1	260.46	15.1	1143.67	523.3

（1）建立1985年我国北方地区的农业产出线性回归模型，并剔除不显著的变量；

（2）试解释说明分析结论。

12. 某地区城镇居民人均全年耐用消费品支出、人均年可支配收入及耐用消费品价格指数的统计资料如表5-15所示。

表5-15　某地居民人均全年耐用消费品支出、人均年可支配收入及耐用消费品价格指数

年份	人均耐用消费品支出 Y/元	人均年可支配收入 X_1/元	耐用消费品价格指数 X_2（1990年为100）
1991	137.16	1181.4	115.96
1992	124.56	1375.7	133.35
1993	107.91	1501.2	128.21
1994	102.96	1700.6	124.85
1995	125.24	2026.6	122.49
1996	162.45	2577.4	129.86
1997	217.43	3496.2	139.52
1998	253.42	4283.0	140.44
1999	251.07	4838.9	139.12
2000	285.85	5160.3	133.35
2001	327.26	5425.1	126.39

利用表中数据，建立该地区城镇居民人均全年耐用消费品支出关于人均年可支配收入和耐用消费品价格指数的回归模型，进行回归分析，并检验人均年可支配收入及耐用消费品价格指数对城镇居民人均全年耐用消费品支出是否有显著影响。

13. 表5-16给出的是1960—1982年间7个OECD国家的能源需求指数（Y）、实际GDP指数（X_1）、能源价格指数（X_2）的数据，所有指数均以1970年为基准（1970年为100）。

表5-16　7个OECD国家的能源指数数据

年份	能源需求指数 Y	实际GDP指数 X_1	能源价格指数 X_2	年份	能源需求指数 Y	实际GDP指数 X_1	能源价格指数 X_2
1960	54.1	54.1	111.9	1972	97.2	94.3	98.6
1961	55.4	56.4	112.4	1973	100.0	100.0	100.0
1962	58.5	59.4	111.1	1974	97.3	101.4	120.1
1963	61.7	62.1	110.2	1975	93.5	100.5	131.0
1964	63.6	65.9	109.0	1976	99.1	105.3	129.6
1965	66.8	69.5	108.3	1977	100.9	109.9	137.7
1966	70.3	73.2	105.3	1978	103.9	114.4	133.7
1967	73.5	75.7	105.4	1979	106.9	118.3	144.5
1968	78.3	79.9	104.3	1980	101.2	119.6	179.0
1969	83.3	83.8	101.7	1981	98.1	121.1	189.4
1970	88.9	86.2	97.7	1982	95.6	120.6	190.9
1971	91.8	89.8	100.3				

（1）建立能源需求与收入和价格之间的对数需求函数 $\ln Y_t=\beta_0+\beta_1\ln X_{1t}+\beta_2\ln X_{2t}+u_t$，解释各回归系数的意义，用 p 值检验所估计回归系数是否显著；

（2）再建立能源需求与收入和价格之间的线性回归模型 $Y_t=\beta_0+\beta_1X_{1t}+\beta_2X_{2t}+u$，解释各回归系数的意义，用 p 值检验所估计回归系数是否显著；

（3）比较所建立的两个模型，如果两个模型结论不同，你将选择哪个模型？为什么？