数学实验

Mathematical Experiments

第13讲

统计方法III　回归分析

实验13　内容提要

回归分析(Regression Analysis)简介

1. 实例及其数学模型

2. 一元线性回归分析

3. 多元线性回归分析

从应用角度介绍回归分析的基本原理、方法和软件实现

回归分析

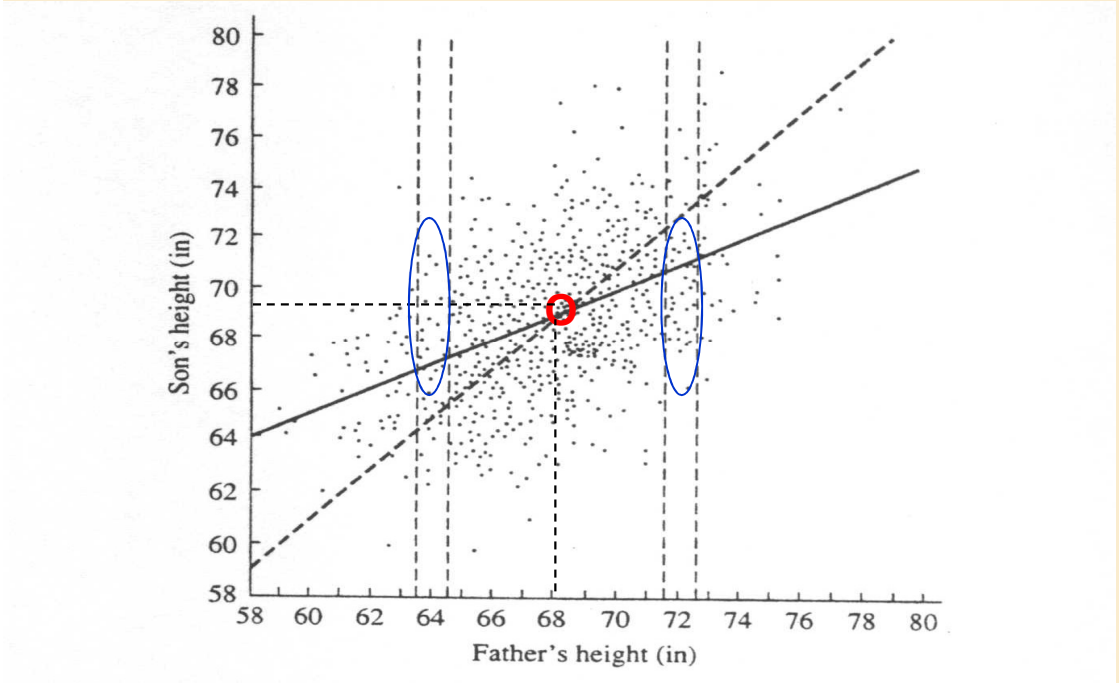
- 回归分析是研究变量间关系的统计学课题
- 在数据的定量分析中，往往需要处理存在着一定联系的变量，需要刻画变量之间有怎样的相互关系，以及如何发生相互影响
- 一元线性回归分析、多元线性回归分析、非线性回归分析、时间序列分析，以及逻辑回归分析等

回归分析的主要步骤

- 收集一组包含因变量和自变量的数据；
- 选定因变量与自变量之间的模型，利用数据按照最小二乘准则计算模型中的系数；
- 给出结果参数的概率解释；
- 判断得到的模型是否适合于这组数据，诊断有无不适合回归模型的异常数据；
- 利用模型对因变量作出预测或解释。

回归(regression) 的由来 Francis Golton (1822-1911)

- 一般说来高个子的父代会有高个子的子代.
- 子代的身高比父代更加趋向一致(“向平庸的回归”) .



$\bar{x} \approx 68, \bar{y} \approx 69$

儿子比父亲平均高**1**英寸

对于身高**72**英寸的父亲，
儿子身高多数不到**73**英寸;

对于身高**64**英寸的父亲，
儿子身高多数超过**65**英寸;

回归直线 $y=0.516\,x+33.73$

Pearson: 1078个父亲和儿子身高的散点图

实例及其数学模型　例1　血压与年龄

为了解血压随年龄增长而升高的关系，调查了30个成年人的血压（收缩压，mmHg）与年龄:

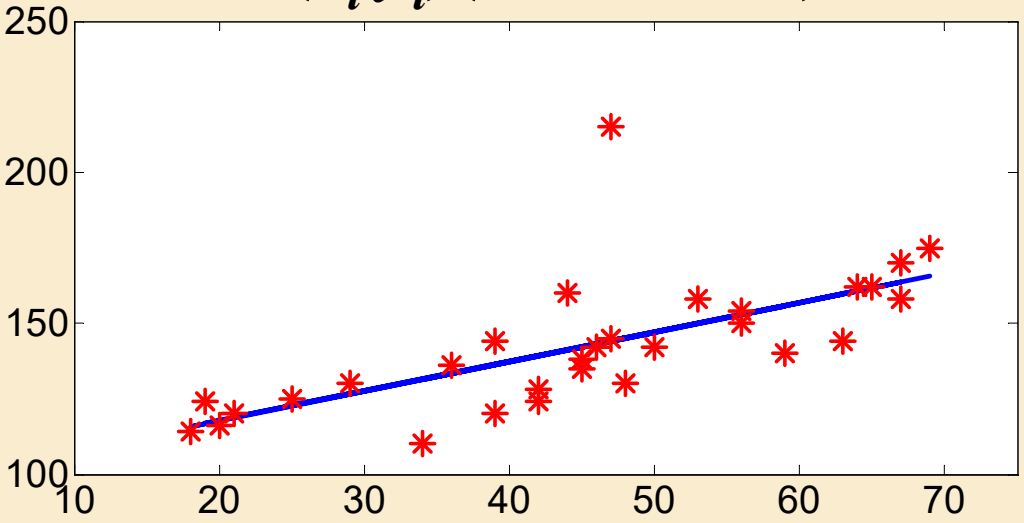
序号	血压	年龄	序号	血压	年龄	序号	血压	年龄
1	144	39	11	162	64	21	136	36
2	215	47	12	150	56	22	142	50
3	138	45	13	140	59	23	120	39
4	145	47	14	110	34	24	120	21
5	162	65	15	128	42	25	160	44
…	…	…	…	…	…	…	…	…

- 用这组数据确定血压与年龄的关系;
- 从年龄预测血压可能的变化范围;
- 回答 “平均说来60岁比50岁的人血压高多少”。

例1 血压与年龄

模型 记血压(因变量) y，年龄(自变量) x，

作数据 $(x_i\ y_i)(i=1,2,\ldots 30)$的散点图



y与x大致呈线性关系

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

由数据确定系数 β_0, β_1

的估计值 $\hat{\beta}_0, \hat{\beta}_1$

- 曲线拟合(求超定线性方程组的最小二乘解);
- 从统计推断角度讨论β_0, β_1 的置信区间和假设检验;
- 对任意的年龄 x 给出血压 y 的预测区间。

例2 血压与年龄、体重指数、吸烟习惯

又调查了例1中30个成年人的体重指数、吸烟习惯:

序号	血压	年龄	体重指数	吸烟	序号	血压	年龄	体重指数	吸烟	序号	血压	年龄	体重指数	吸烟
1	144	39	24.2	0	11	162	64	28.0	1	21	136	36	25.0	0
2	215	47	31.1	1	12	150	56	25.8	0	22	142	50	26.2	1
3	138	45	22.6	0	13	140	59	27.3	0	23	120	39	23.5	0
4	145	47	24.0	1	14	110	34	20.1	0	24	120	21	20.3	0
5	162	65	25.9	1	15	128	42	21.7	0	25	160	44	27.1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

体重指数: 体重(kg) /[身高(m)]^2

吸烟习惯: 0~不吸烟，1~吸烟

例2 血压与年龄、体重指数、吸烟习惯

模型　记血压 y，年龄x_1、体重指数x_2、吸烟习惯x_3

作数据 y 对x_2 的散点图　　y与x_2大致呈线性关系

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

由数据确定系数 $\beta_0, \beta_1, \beta_2, \beta_3$

的估计值 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

例3　软件开发人员的薪金

建立模型研究薪金与资历、管理责任、教育程度的关系，分析人事策略的合理性，作为新聘用人员薪金的参考.

46名软件开发人员的档案资料

编号	薪金	资历	管理	教育	编号	薪金	资历	管理	教育
01	13876	1	1	1	42	27837	16	1	2
02	11608	1	0	3	43	18838	16	0	2
03	18701	1	1	3	44	17483	16	0	1
04	11283	1	0	2	45	19207	17	0	2
05	11767	1	0	3	46	19346	20	0	1

资历~ 从事专业工作的年数；管理~ 1=管理人员，0=非管理人员；教育~ 1=中学，2=大学，3=研究生

模型　　y~ 薪金，x_1~资历（年）

$x_2=1$~ 管理人员，$x_2=0$~ 非管理人员

教育　1=中学　　$x_3=\begin{cases}1, & \text{中学}\\ 0, & \text{其它}\end{cases}$

2=大学

3=研究生　　$x_4=\begin{cases}1, & \text{大学}\\ 0, & \text{其它}\end{cases}$

⇨ 中学：$x_3=1, x_4=0$；

大学：$x_3=0, x_4=1$；

研究生：$x_3=0, x_4=0$

假设

- 资历每加一年薪金的增长是常数；
- 管理、教育、资历之间无交互作用.

线性回归模型　$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$

由数据确定 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

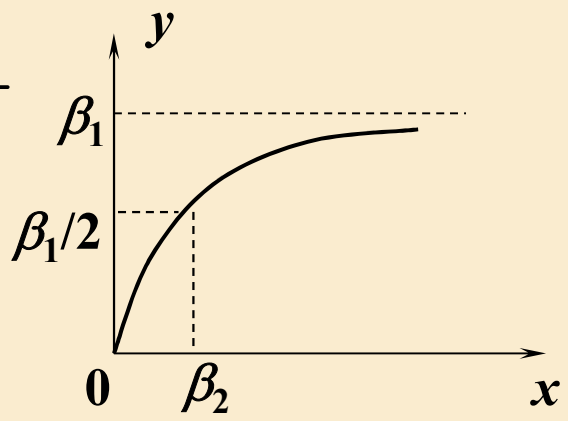
例4　酶促反应

酶~高效生物催化剂; 酶促反应~经过酶催化的化学反应

酶促反应的反应速度主要取决于反应物（底物）的浓度:

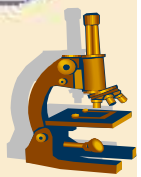
- 底物浓度较小时，反应速度大致与浓度成正比;
- 底物浓度很大、渐进饱和时，反应速度趋于固定值.

Michaelis-Menten模型　$y = \dfrac{\beta_1 x}{\beta_2 + x}$



y ~ 酶促反应的速度, x ~底物浓度

待定系数 β_1 (最终反应速度)　　β_2 (半速度点)



例4　酶促反应

为研究酶促反应中嘌呤霉素对反应速度与底物浓度之间关系的影响，设计了两个实验：使用的酶经过嘌呤霉素处理；使用的酶未经嘌呤霉素处理。

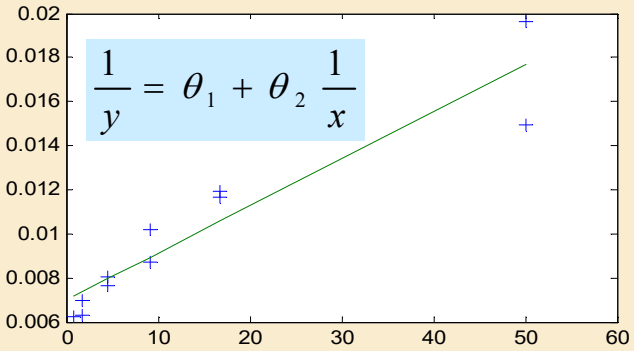
实验数据

底物浓度(ppm)		0.02		0.06		0.11		0.22		0.56		1.10	
反应速度	处理	76	47	97	107	123	139	159	152	191	201	207	200
	未处理	67	51	84	86	98	115	131	124	144	158	160	/

对未经嘌呤霉素处理的反应，用实验数据估计参数β_1, β_2；用实验数据研究嘌呤霉素处理对参数β_1, β_2的影响。

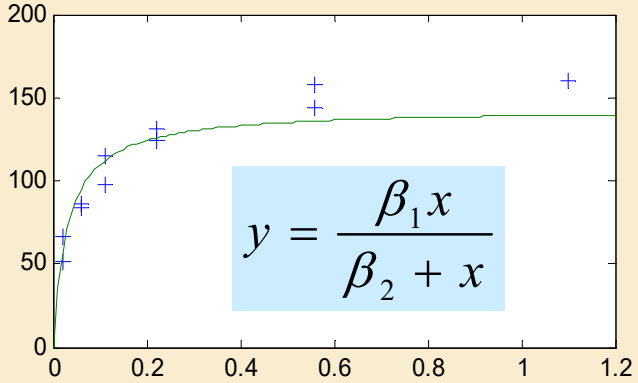
模型 $y=\dfrac{\beta_1 x}{\beta_2+x}$ ⇨ $\dfrac{1}{y}=\dfrac{1}{\beta_1}+\dfrac{\beta_2}{\beta_1}\dfrac{1}{x}=\theta_1+\theta_2\dfrac{1}{x}$

对β_1, β_2非线性　　对θ_1, θ_2线性



1/*x*较小时有很好的线性趋势，
1/*x*较大时出现很大的分散.

$\theta_1=6.972\times10^{-3}$, $\theta_2=0.215\times10^{-3}$ ⇨ $\beta_1=143.43$, $\beta_2=0.0308$



***x*较大时，*y*有较大偏差.**
参数估计时，*x*较小（1/*x*很大）的数据控制了参数的确定.

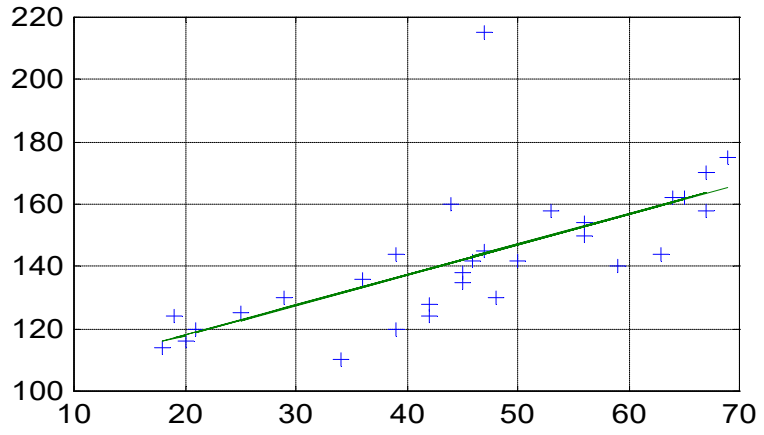
直接考虑非线性模型

一元线性回归分析

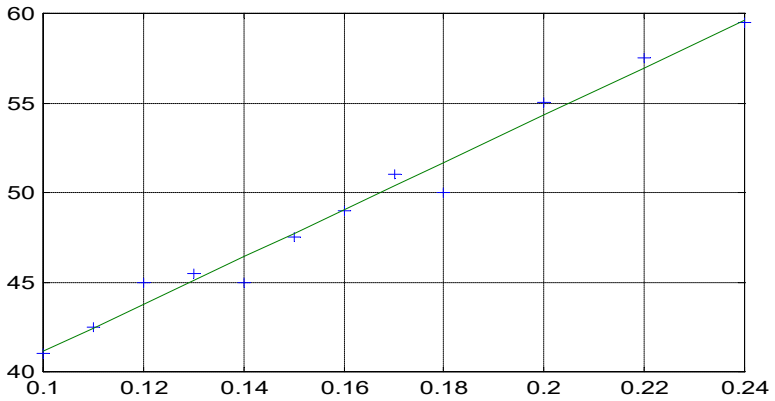
问题 已知一组数据 (x_i, y_i), $i=1,2,\ldots n$ (平面上的n个点)，

用最小二乘准则确定一个线性函数(直线) $y = \hat{\beta}_0 + \hat{\beta}_1 x$

1. 血压与年龄



2. 合金强度与碳含量



怎样衡量由最小二乘准则拟合得到的模型的可靠程度？

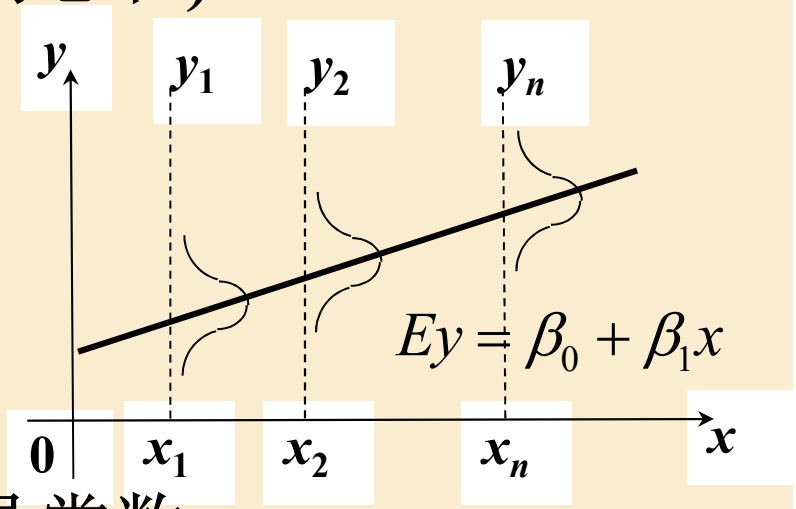
怎样给出模型系数的置信区间和因变量的预测区间？

一元线性回归模型

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

x~自变量　　β_0，β_1 ~回归系数

ε~随机变量(影响y的随机因素的总和)



基本假设

独立性：x_i**相互独立**，y_i**相互独立**

线性性：y的期望是x的线性函数

齐次性：对于不同的x，y的方差是常数

正态性：对于给定的x，y服从正态分布

ε是相互独立的、期望为0、方差为σ^2、正态分布的随机变量，即$\varepsilon \sim N(0, \sigma^2)$，$\varepsilon$称（随机）误差。

一元线性回归的方差分析

偏差的分解：$y_i-\bar{y}=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})$

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

$$S \quad = \quad U \quad + \quad Q$$

总偏差平方和　　回归平方和　　残差平方和

决定系数　$R^2=U/S$

因变量的总变化中自变量引起的部分的比例

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i \qquad \bar{y}=\hat{\beta}_0+\hat{\beta}_1\bar{x}$$

回归方程的显著性检验，F检验

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

$$(x_i,y_i)\Rightarrow \hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$$

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x+\varepsilon)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

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

$$\frac{Q}{\sigma^2}\sim\chi^2(n-2)$$

$$\hat{\beta}_1\sim N\left(\beta_1,\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

$$\hat{\beta}_0\sim N\left(\beta_0,\ \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right)$$

回归方程的显著性检验，F检验

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

当$H_0:\beta_1=0$成立时，$\hat{\beta}_1\sim N\left(0,\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\Rightarrow\dfrac{\hat{\beta}_1\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\sigma}\sim N(0,1)$

回归平方和 $U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2\Rightarrow\dfrac{U}{\sigma^2}\sim\chi^2(1)$

且U与Q独立。 $F=\dfrac{U}{Q/(n-2)}\sim F(1,n-2)$

回归方程系数的 t 检验和区间估计

$$\hat{\beta}_0 \sim N\left(\beta_0,\ \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right),\quad \hat{\beta}_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

$$\frac{Q}{\sigma^2} \sim \chi^2(n-2)$$

$$S_{\beta_0}{}^2=\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\cdot\frac{Q}{n-2},\qquad S_{\beta_1}{}^2=\frac{Q}{(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

当$H_0:\beta_0=0$成立时，$\dfrac{\hat{\beta}_0}{S_{\beta_0}} \sim t(n-2)$；　当$H_0:\beta_1=0$成立时，$\dfrac{\hat{\beta}_1}{S_{\beta_1}} \sim t(n-2)$

β_0的区间估计 $\left[\hat{\beta}_0-t_{1-\alpha/2}(n-2)\cdot S_{\beta_0},\ \hat{\beta}_0+t_{1-\alpha/2}(n-2)\cdot S_{\beta_0}\right]$

β_1的区间估计 $\left[\hat{\beta}_1-t_{1-\alpha/2}(n-2)\cdot S_{\beta_1},\ \hat{\beta}_1+t_{1-\alpha/2}(n-2)\cdot S_{\beta_1}\right]$

一元线性回归的MATLAB实现

b=regress(y,X)

[b,bint,r,rint,s]=regress(y,X,alpha)

输入：**y~**因变量（列向量），**X~1**与自变量组成的矩阵，
alpha~显著性水平α（缺省时设定为**0.05**）.

输出：$b=(\hat{\beta}_0,\hat{\beta}_1)$，**bint~**$\beta_0$, β_1的置信区间，
r~残差（列向量），**rint~**残差的置信区间.

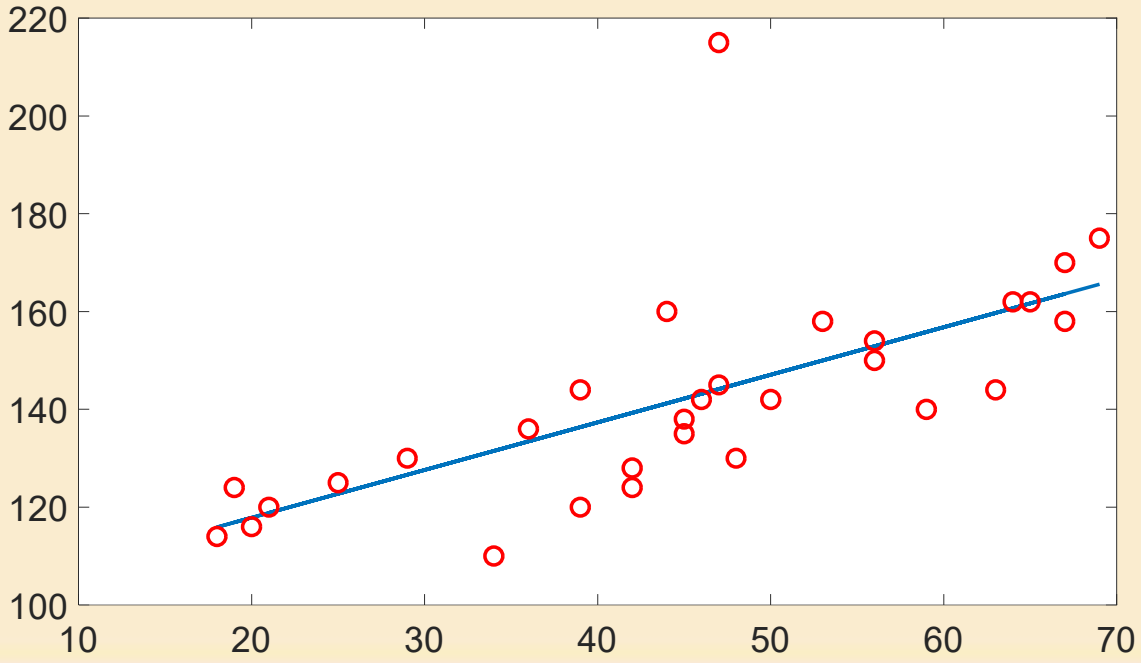
s(4个统计量**)**：
决定系数R^2；F值；$F(1,n-2)$分布大于F值的概率p；剩余方差s^2.
当 $p<\alpha$ 时拒绝H_0，回归模型有效.

```
y=[144  215 138 145 162 142 170 124 158 154 162 150 140 110
128 130 135 114 116 124 136 142 120 120 160 158 144 130 125
175];
x=[39   47  45  47  65  46  67  42  67  56 64   56  59  34
42  48  45  18  20  19 36   50  39  21  44  53  63  29  25
69];
n=length(y);
X=[ones(n,1)  x'];
[b,bint,r,rint,s]=regress(y',X);
b,bint,s,s2=sum(r.^2)/(n-2)
pause
plot(x,b(1)+b(2)*x,'LineWidth',2);   % 最小二乘拟合直线
hold on
plot(x,y,'ro','MarkerSize',10,'LineWidth',2); 散点图
```

例1 血压与年龄 模型 $y = \beta_0 + \beta_1 x$ 数据

回归系数	回归系数估计值	回归系数置信区间
β_0	98.4084	[78.7484 118.0683]
β_1	0.9732	[0.5601 1.3864]
R^2=0.4540, F=23.2834, p<0.0001, s^2 = 273.7137		

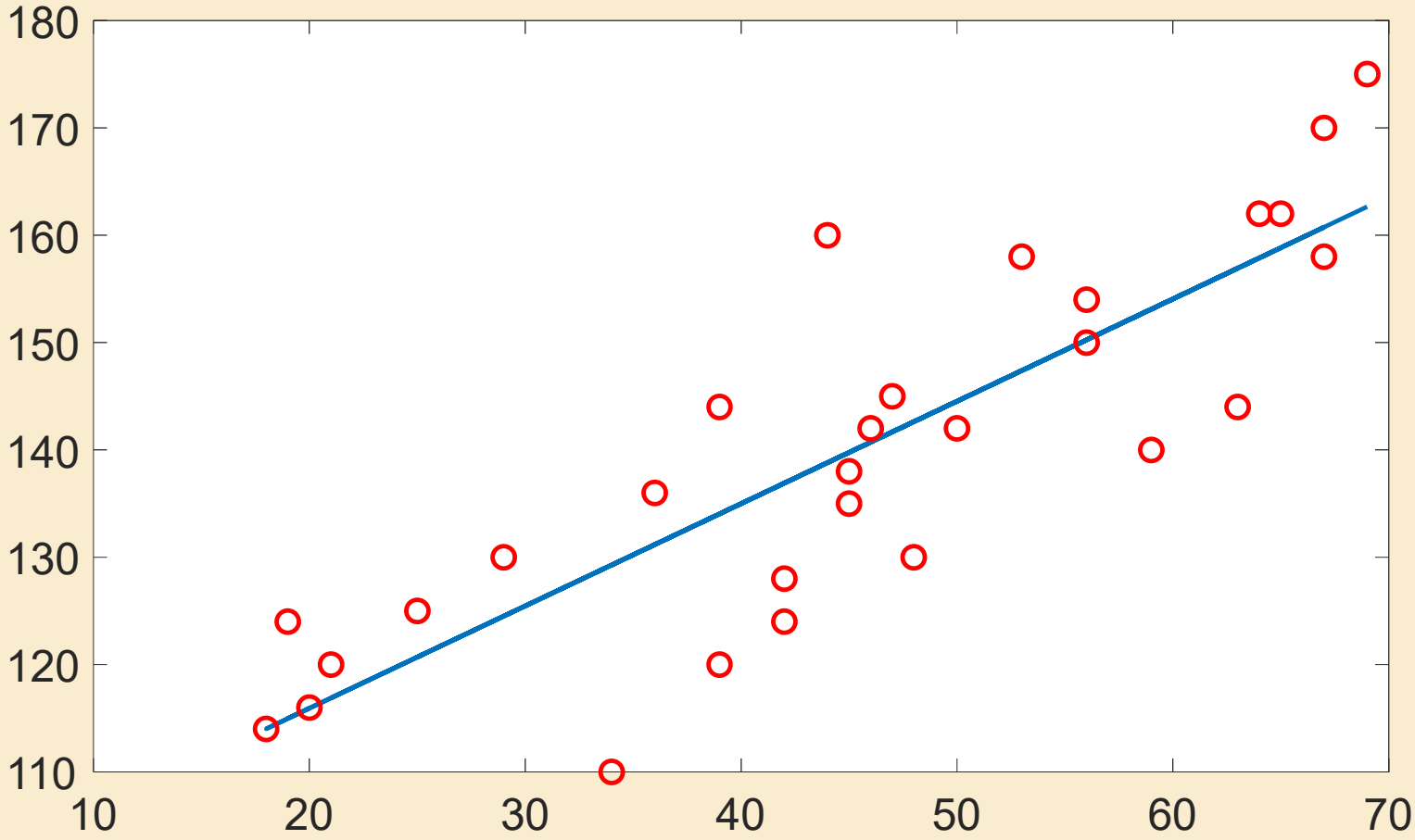
s2=sum(r.^2)/(n-2) =273.7137

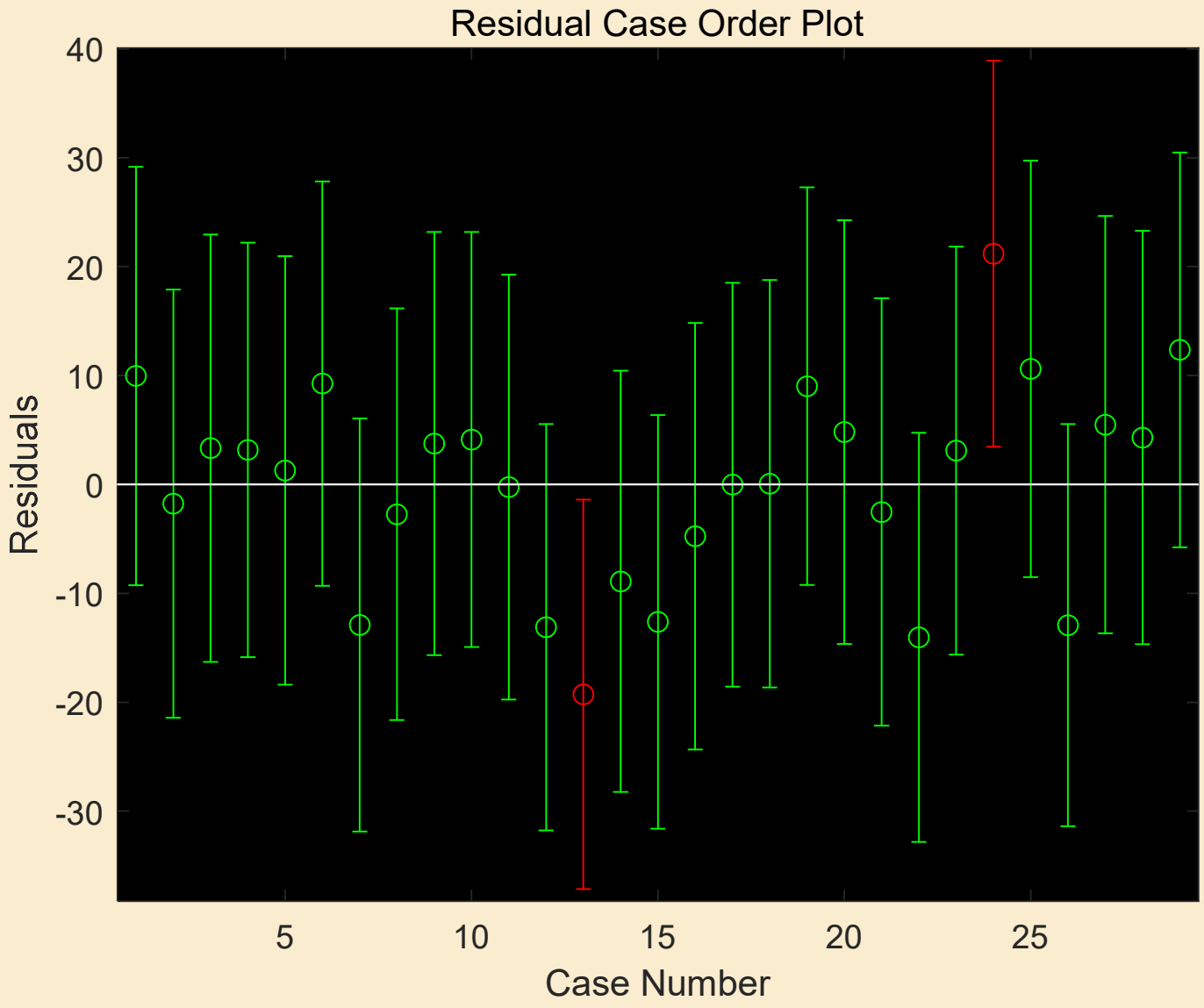


- `close; rcoplot(r,rint)` %残差分析图

```
y(2)=[]; x(2)=[];   % 去除异常点
n=length(y); X=[ones(n,1)  x'];
[b,bint,r,rint,s]=regress(y',X);
plot(x,b(1)+b(2)*x,'LineWidth',2);   % 最小二乘拟合直线
hold on
plot(x,y,'ro','MarkerSize',10,'LineWidth',2); % 散点图
pause; close; rcoplot(r,rint)
```

回归系数	回归系数估计值	回归系数置信区间
β_0	96.8665	[85.4771 108.2559]
β_1	0.9533	[0.7140 1.1925]
R^2= 0.7123, F= 66.8358, p<0.0001, s^2 =91.4305		


180
170
160
150
140
130
120
110
10
20
30
40
50
60
70


Residual Case Order Plot
Residuals
40
30
20
10
0
-10
-20
-30
5
10
15
20
25
Case Number

利用一元线性回归模型进行预测

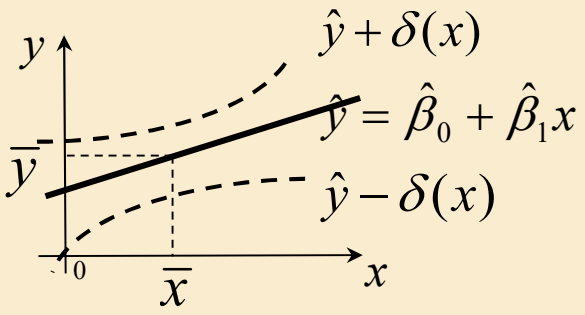
x_0给定，y_0的预测值：　$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

性质：$\hat{y}_0$ 无偏，且 $E(\hat{y}_0 - y_0)^2$ 最小　　$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

预测区间

$$\left[\hat{y}_0 - t_{1-\alpha/2}(n-2)s\sqrt{\frac{(x_0-\bar{x})^2}{s_{xx}} + \frac{1}{n} + 1},\ \hat{y}_0 + t_{1-\alpha/2}(n-2)s\sqrt{\frac{(x_0-\bar{x})^2}{s_{xx}} + \frac{1}{n} + 1} \right]$$

s~剩余标准差



n很大且 x_0接近 $\bar{x}$

$$[\hat{y}_0 - u_{1-\alpha/2}s,\ \hat{y}_0 + u_{1-\alpha/2}s]$$

$$\delta(x) = t_{(n-2),1-\alpha/2}s\sqrt{\frac{(x-\bar{x})^2}{s_{xx}} + \frac{1}{n} + 1} \approx u_{1-\alpha/2}s$$

预测区间

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0 \sim N\left(\beta_0+\beta_1 x_0, \sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{XX}}\right]\right)$$

$$y_0=\beta_0+\beta_1 x_0+\varepsilon_0 \sim N\left(\beta_0+\beta_1 x_0, \sigma^2\right) \Rightarrow Var\left(y_0-\hat{y}_0\right)=\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{XX}}\right]$$

$$y_0-\hat{y}_0 \sim N\left(0, \sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{XX}}\right]\right)$$

$$\frac{Q}{\sigma^2} \sim \chi^2(n-2)$$

$$T=\frac{\dfrac{y_0-\hat{y}_0}{\sigma\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{s_{XX}}}}}{\sqrt{\dfrac{Q}{\sigma^2(n-2)}}}=\frac{y_0-\hat{y}_0}{\sqrt{\dfrac{Q}{(n-2)}}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{s_{XX}}}}=\frac{y_0-\hat{y}_0}{\hat{\sigma}_e\cdot\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{s_{XX}}}}=\sim t(n-2)$$

```
x0=50;
y0=b(1)+b(2)*x0;                    % 预测y(x=50)
xb=mean(x);
sxx=sum((x-xb).^2);
a=sqrt((x0-xb)^2/sxx+1/n+1);
t=tinv(0.975,n-2);
d=t*a*sqrt(s2);
y1=y0-d;y2=y0+d;                    %预测y(x=50)区间（t分布）
[y0 y1 y2]
d1=norminv(0.975)*sqrt(s2);
y3=y0-d1;y4=y0+d1;
[y0 y3 y4]                          %预测y(x=50)区间（正态分布）
```

144.5298 124.5406 164.5190

144.5298 125.7887 163.2708

多元线性回归分析

模型 $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$

估计回归系数 ⇩ $(y_i,x_{i1},\cdots x_{im}),\ i=1,\cdots n,\ n>m$

$$y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_mx_{im}+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma^2),\ i=1,\cdots n$$

⇩

$$X=\begin{bmatrix}1 & x_{11} & \cdots & x_{1m}\\ \cdots & & & \\ 1 & x_{n1} & \cdots & x_{nm}\end{bmatrix},\quad Y=\begin{bmatrix}y_1\\ \cdots\\ y_n\end{bmatrix},\quad \varepsilon=\begin{bmatrix}\varepsilon_1\\ \cdots\\ \varepsilon_n\end{bmatrix},\quad \beta=[\beta_0,\beta_1,\cdots\beta_m]^T$$

$$\begin{cases}Y=X\beta+\varepsilon\\ \varepsilon\sim N(0,\sigma^2)\end{cases}$$

多元线性回归的统计分析

1.误差方差σ^2的估计

	一元回归	多元回归
模型	$y=\beta_0+\beta_1 x_1+\varepsilon$	$y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\varepsilon$
估计值	$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$	$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\cdots+\hat{\beta}_m x_{mi}$
残差	$e_i=\hat{\varepsilon}_i=y_i-\hat{y}_i$	$e_i=\hat{\varepsilon}_i=y_i-\hat{y}_i$
残差平方和	$Q=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$Q=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
剩余方差	$s^2=\hat{\sigma}^2=\frac{Q}{n-2}$	$s^2=\hat{\sigma}^2=\frac{Q}{n-m-1}$
Q的自由度	***n*-2 (2个参数)**	***n*-(*m*+1) (*m*+1个参数)**

2. 回归系数的区间估计和假设检验

一元回归

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / s_{xx}), \quad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$Q/\sigma^2 \sim \chi^2(n-2),$$

$$t = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{s_{xx}}/\sigma}{\sqrt{Q/(n-2)\sigma^2}} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{s_{xx}}}{s} \sim t(n-2)$$

$$[\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2)\frac{s}{\sqrt{s_{xx}}}]$$

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

$$|t| = \left|\frac{\hat{\beta}_1\sqrt{s_{xx}}}{s}\right| > t_{1-\alpha/2}(n-2)$$

多元回归

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj}), \quad c_{jj} \sim (\tilde{X}^T \tilde{X})^{-1}$$

的j对角元

$$Q/\sigma^2 \sim \chi^2(n-m-1)$$

$$t_j = \frac{(\hat{\beta}_j - \beta_j)/\sigma\sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)\sigma^2}} = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t(n-m-1)$$

$$[\hat{\beta}_j \pm t_{1-\alpha/2}(n-m-1)s\sqrt{c_{jj}}]$$

$$H_0^{(j)} : \beta_j = 0, \quad H_1^{(j)} : \beta_j \neq 0$$

$$|t_j| = \left|\frac{\hat{\beta}_j}{s\sqrt{c_{jj}}}\right| > t_{1-\alpha/2}(n-m-1)$$

拒绝H_0, 模型有效

3. 模型的有效性检验

	一元回归	多元回归
偏差分解	$S = U + Q$	$S = U + Q$
决定系数	$R^2 = U/S$	$R^2 = U/S$
	$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$	$H_0^{(j)} : \beta_j = 0, \quad H_1^{(j)} : \beta_j \neq 0$
H_0成立	$U/\sigma^2 \sim \chi^2(1), \quad Q/\sigma^2 \sim \chi^2(n-2),$	$U/\sigma^2 \sim \chi^2(m), \; Q/\sigma^2 \sim \chi^2(n-m-1)$
	$F = \dfrac{U}{Q/(n-2)} \sim F(1, n-2)$	$F = \dfrac{U/m}{Q/(n-m-1)} \sim F(m, n-m-1)$
检验	$F > F_{1-\alpha}(1, n-2)$	$F > F_{1-\alpha}(m, n-m-1)$

拒绝H_0, 模型有效

利用多元线性回归模型进行预测

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$$

性质：$\hat{y}_0$ 无偏，且 $E(\hat{y}_0 - y_0)^2$ 最小

预测区间 $[\hat{y} - \delta(x),\ \hat{y} + \delta(x)]$

$$\delta(x) = t_{1-\alpha/2}(n-m-1)\cdot s\sqrt{(x-\bar{x})^T(\tilde{X}^T\tilde{X})^{-1}(x-\bar{x}) + \frac{1}{n} + 1} \approx u_{1-\alpha/2}s$$

与一元回归对比 $\delta(x) = t_{1-\alpha/2}(n-2)\cdot s\sqrt{\frac{(x-\bar{x})^2}{s_{xx}} + \frac{1}{n} + 1} \approx u_{1-\alpha/2}s$

多元线性回归的MATLAB实现

与一元回归相同

b=regress(y,X) 注意 X 的构造

[b,bint,r,rint,s]=regress(y,X,alpha)

例2 血压与年龄、体重指数、吸烟习惯 Exp13.m

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$ 剔除两个异常点后

$$\hat{y} = 58.5101 + 0.4303 x_1 + 2.3449 x_2 + 10.3065 x_3$$

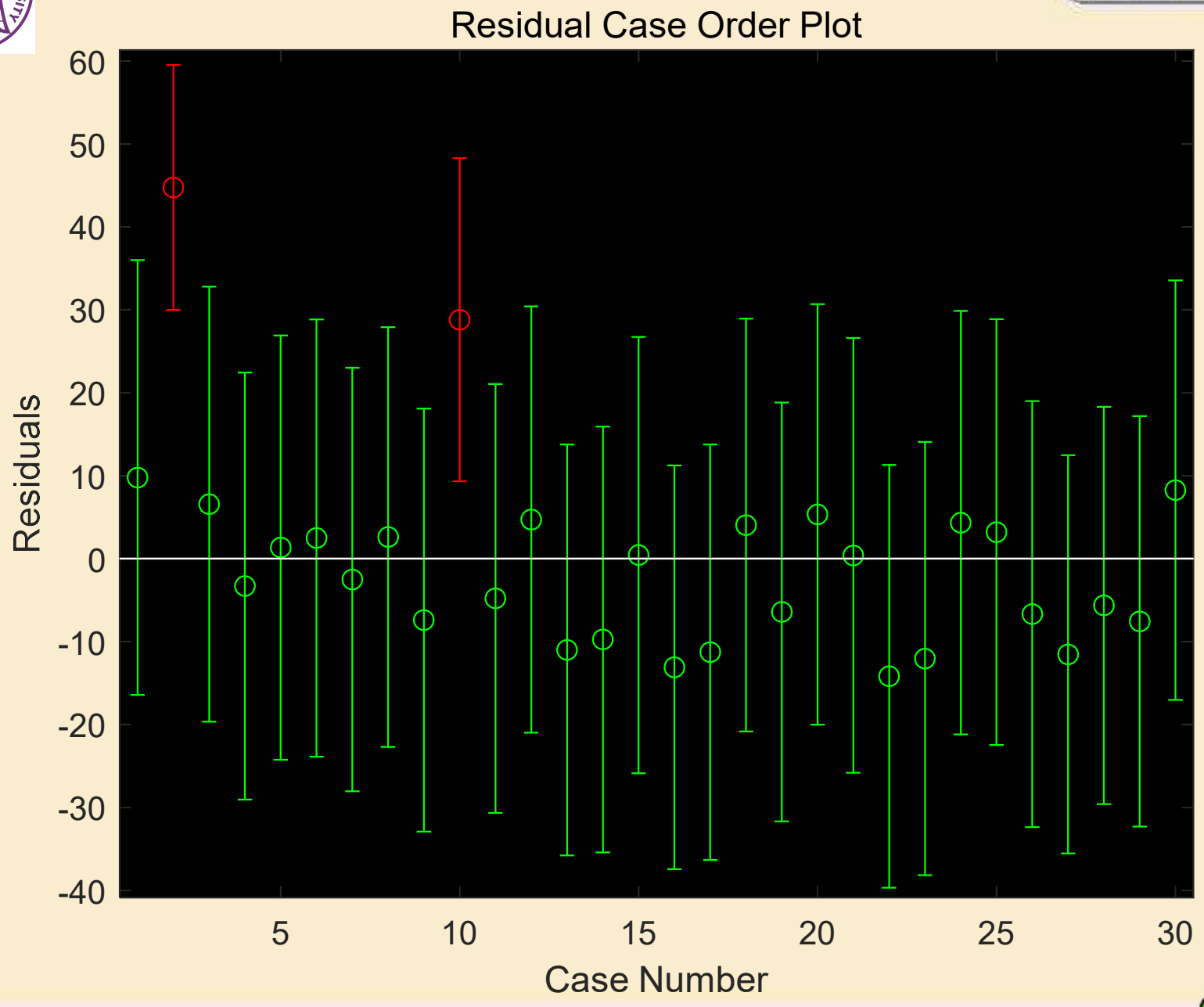
- 年龄和体重指数相同，吸烟者比不吸烟者的血压(平均)高**10. 3**
- 与例**1** “血压与年龄”的结果 $\hat{y} = 96.8665 + 0.9533 x_1$ 相比，

  年龄增加**1**岁血压的升高值(即β_1)为何有这么大的差别**?**

- b =
- 45.3636
- 0.3604
- 3.0906
- 11.8246

- bint =
- 3.5537 87.1736
- -0.0758 0.7965
- 1.0530 5.1281
- -0.1482 23.7973

- s =
- 0.6855 18.8906 0.0000 169.7917


Residual Case Order Plot
Residuals
60
50
40
30
20
10
0
-10
-20
-30
-40
5
10
15
20
25
30
Case Number

- b =
- 58.5101
- 0.4303
- 2.3449
- 10.3065

- bint =
- 29.9064 87.1138
- 0.1273 0.7332
- 0.8509 3.8389
- 3.3878 17.2253

- s =
- 0.8462 44.0087 0.0000 53.6604


Residual Case Order Plot
Residuals
25
20
15
10
5
0
-5
-10
-15
-20
-25
5
10
15
20
25
Case Number

线性最小二乘拟合与多元线性回归的一般形式

线性回归模型 $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$ (1)

“线性”是指**y**是系数β的关系(非指**y**与**x**的关系)

$y=\beta_0+\beta_1x^2,\quad y=\beta_0+\beta_1e^{x_1}+\beta_2/x_2$ ~ 线性回归

线性回归一般形式 $y=\beta_0+\beta_1r_1(x)+\cdots+\beta_mr_m(x)+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$ (2)

$x=(x_1,\cdots x_k),\quad r_j(x)\ (j=1,\cdots,m)$ 是已知函数

令$r_j(x)=u_j$, 则(2)→(1)

多元线性回归中的交互作用

例3 软件开发人员的薪金 $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$

y~薪金, x_1~资历, $x_2=1$~ 管理人员, $x_2=0$~ 非管理人员

x_3=1, x_4=0~中学； x_3=0, x_4=1~大学； x_3=0, x_4=0~研究生

系数	系数估计	置信区间
β_0	11033	[10258 11807]
β_1	546	[484 608]
β_2	6883	[6248 7517]
β_3	-2994	[-3826 -2162]
β_4	148	[-636 931]
R^2=0.957	F=226	p<0.0001

R^2,F, $p\rightarrow$ 模型整体上可用

资历增加1年
薪金增长546

管理人员多6883

中学程度比更高的少2994

大学程度比更高的多148

β_4置信区间包含零点，
解释不可靠！

用残差分析发现交互作用

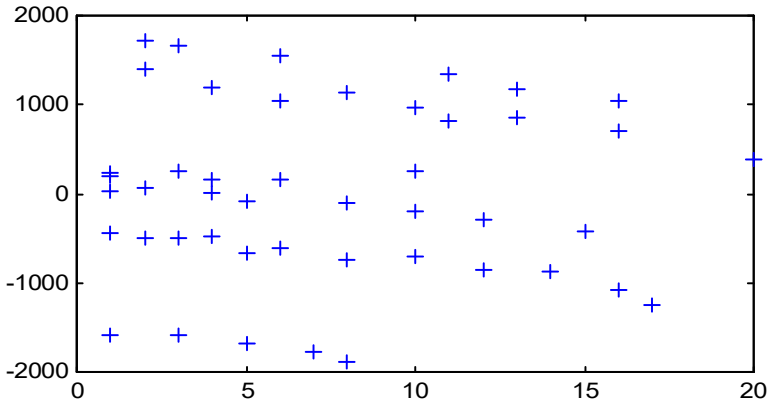
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

考察残差 $e = y - \hat{y}$ 是否为$N(0, \sigma^2)$

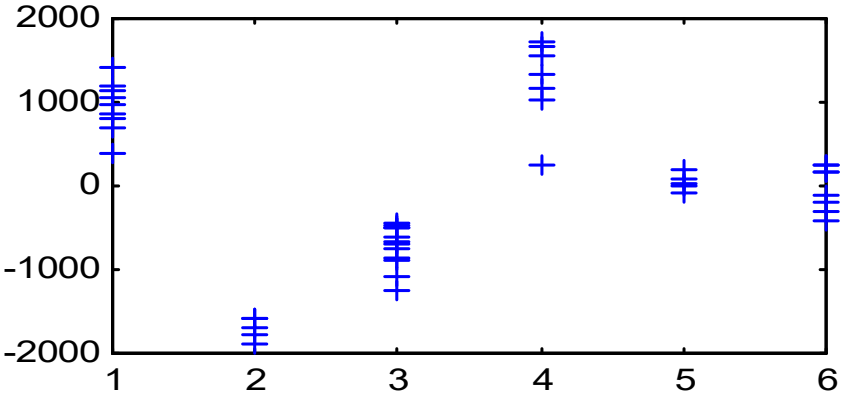
管理与教育的组合

组合	1	2	3	4	5	6
管理	0	1	0	1	0	1
教育	1	1	2	2	3	3

e 与资历x_1的关系



e与管理—教育组合的关系



残差大概分成3个水平，6种管理—教育组合混在一起，未正确反映

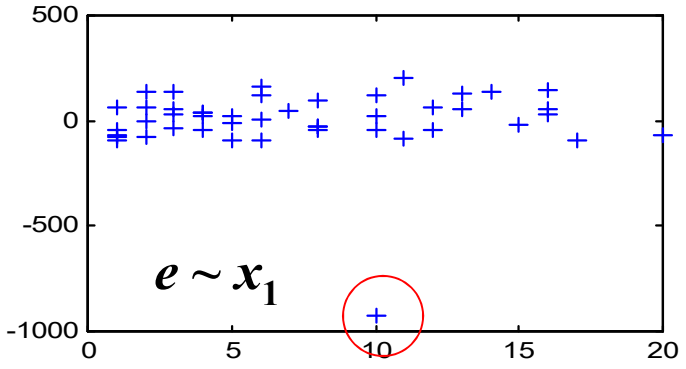
残差全为正，或全为负，管理—教育组合处理不当

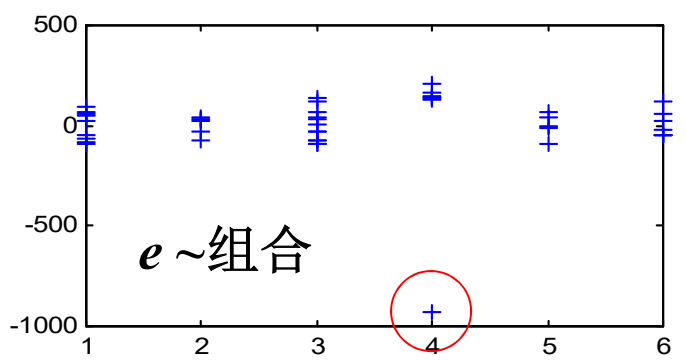
应增加x_2与x_3, x_4的交互项

增加管理x_2与教育x_3, x_4的交互项

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_2x_3+\beta_6x_2x_4+\varepsilon$$

系数	系数估计值	置信区间
β_0	11204	[11044 11363]
β_1	497	[486 508]
β_2	7048	[6841 7255]
β_3	-1727	[-1939 -1514]
β_4	-348	[-545 –152]
β_5	-3071	[-3372 -2769]
β_6	1836	[1571 2101]
R^2=0.999	F=554	p<0.0001





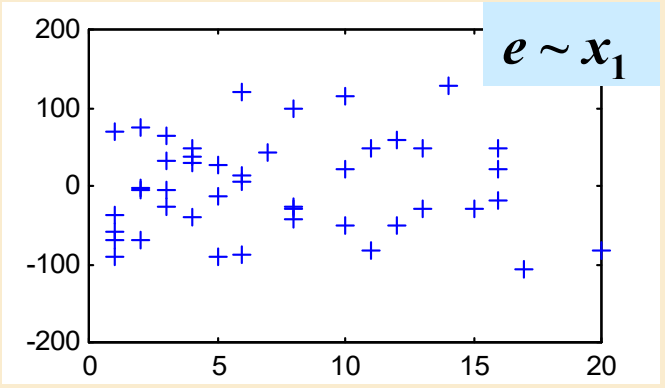
R^2, F有改进，所有回归系数置信区间都不含零点，模型完全可用

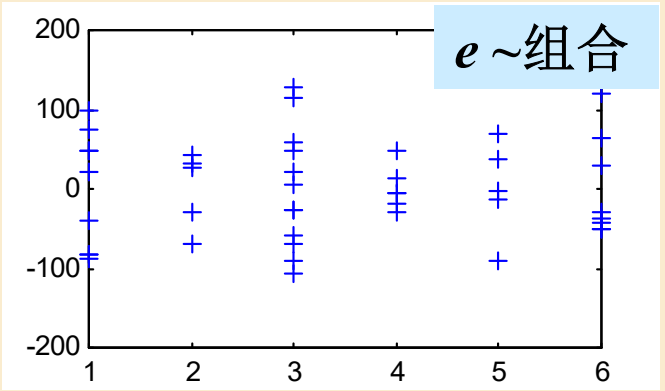
消除了不正常现象

异常数据(33号)应去掉

去掉异常数据后的结果

系数	系数估计值	置信区间
β_0	11200	[11139 11261]
β_1	498	[494 503]
β_2	7041	[6962 7120]
β_3	-1737	[-1818 -1656]
β_4	-356	[-431 –281]
β_5	-3056	[-3171 –2942]
β_6	1997	[1894 2100]
R^2= 0.9998	F=36701	p<0.0001





R^2: 0.957 → 0.999 → 0.9998

F: 226 → 554 → 36701

置信区间长度更短

残差图十分正常

最终模型的结果可以应用

模型应用

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3+\hat{\beta}_4x_4+\hat{\beta}_5x_2x_3+\hat{\beta}_6x_2x_4$$

制订6种管理—教育组合人员的“基础”薪金(资历x_1=0)

组合	管理x_2	教育 (x_3, x_4)	系数	“基础”薪金
1	0	(1,0)	$\beta_0+\beta_3$	9463
2	1	(1,0)	$\beta_0+\beta_2+\beta_3+\beta_5$	13448
3	0	(0,1)	$\beta_0+\beta_4$	10844
4	1	(0,1)	$\beta_0+\beta_2+\beta_4+\beta_6$	19882
5	0	(0,0)	β_0	11200
6	1	(0,0)	$\beta_0+\beta_2$	18241

大学程度管理人员比更高程度管理人员的薪金高

大学程度非管理人员比更高程度非管理人员的薪金略低

实 验 练 习

目的 1、了解回归分析的基本原理

2、根据问题的要求提出模型

3、对已确定的模型，使用MATLAB确定参数

作业 4.（睡眠），　5.(犯罪)