第2章 一元线性回归

2.1 一元线性回归模型
2.2 参数β_0、β_1的估计
2.3 最小二乘估计的性质
2.4 回归方程的显著性检验
2.5 残差分析
2.6 回归系数的区间估计
2.7 预测和控制
2.8 本章小结与评注

2.1 一元线性回归模型

例2.1 表2.1列出了15起火灾事故的损失及火灾发生地与最近的消防站的距离。

表2.1　　火灾损失表

距消防站离 x(km)	3．4	1．8	4．6	2．3	3．1	5．5	0．7	3．0
火灾损失 y(千元)	26.2	17.8	31.3	23.1	27.5	36.0	14.1	22.3
距消防站离 x(km)	2．6	4．3	2．1	1．1	6．1	4．8	3．8	
火灾损失 y(千元)	19.6	31.3	24.0	17.3	43.2	36.4	26.1	

2.1 一元线性回归模型

例2.2 全国人均消费金额记作y(元);
人均国民收入记为x(元)

表2.2　　人均国民收入表

年份	人均国民收入（元）	人均消费金额（元）	年份	人均国民收入（元）	人均消费金额（元）
1980	460	234.75	1990	1634	797.08
1981	489	259.26	1991	1879	890.66
1982	525	280.58	1992	2287	1063.39
1983	580	305.97	1993	2939	1323.22
1984	692	347.15	1994	3923	1736.32
1985	853	433.53	1995	4854	2224.59
1986	956	481.36	1996	5576	2627.06
1987	1104	545.40	1997	6053	2819.36
1988	1355	687.51	1998	6392	2958.18
1989	1512	756.27			

2.1 一元线性回归模型

一元线性回归模型 $y=\beta_0+\beta_1 x+\varepsilon$

$$\begin{cases} E(\varepsilon)=0 \\ \mathrm{var}(\varepsilon)=\sigma^2 \end{cases}$$

回归方程 $E(y|x)=\beta_0+\beta_1 x$

2.1 一元线性回归模型

样本观测值$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$

样本模型 $y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\qquad i=1,2,\ldots,n$

$$\begin{cases} E(\varepsilon_i)=0 \\ \mathrm{var}(\varepsilon_i)=\sigma^2 \end{cases} \qquad i=1,2,\cdots,n$$

回归方程 $E(y_i)=\beta_0+\beta_1 x_i$,$\mathrm{var}(y_i)=\sigma^2$,

经验回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$

2.2 参数β_0、β_1的估计

一、普通最小二乘估计

(Ordinary Least Square Estimation,简记为OLSE)

最小二乘法就是寻找参数β_0、β_1的估计值使离差平方和达极小

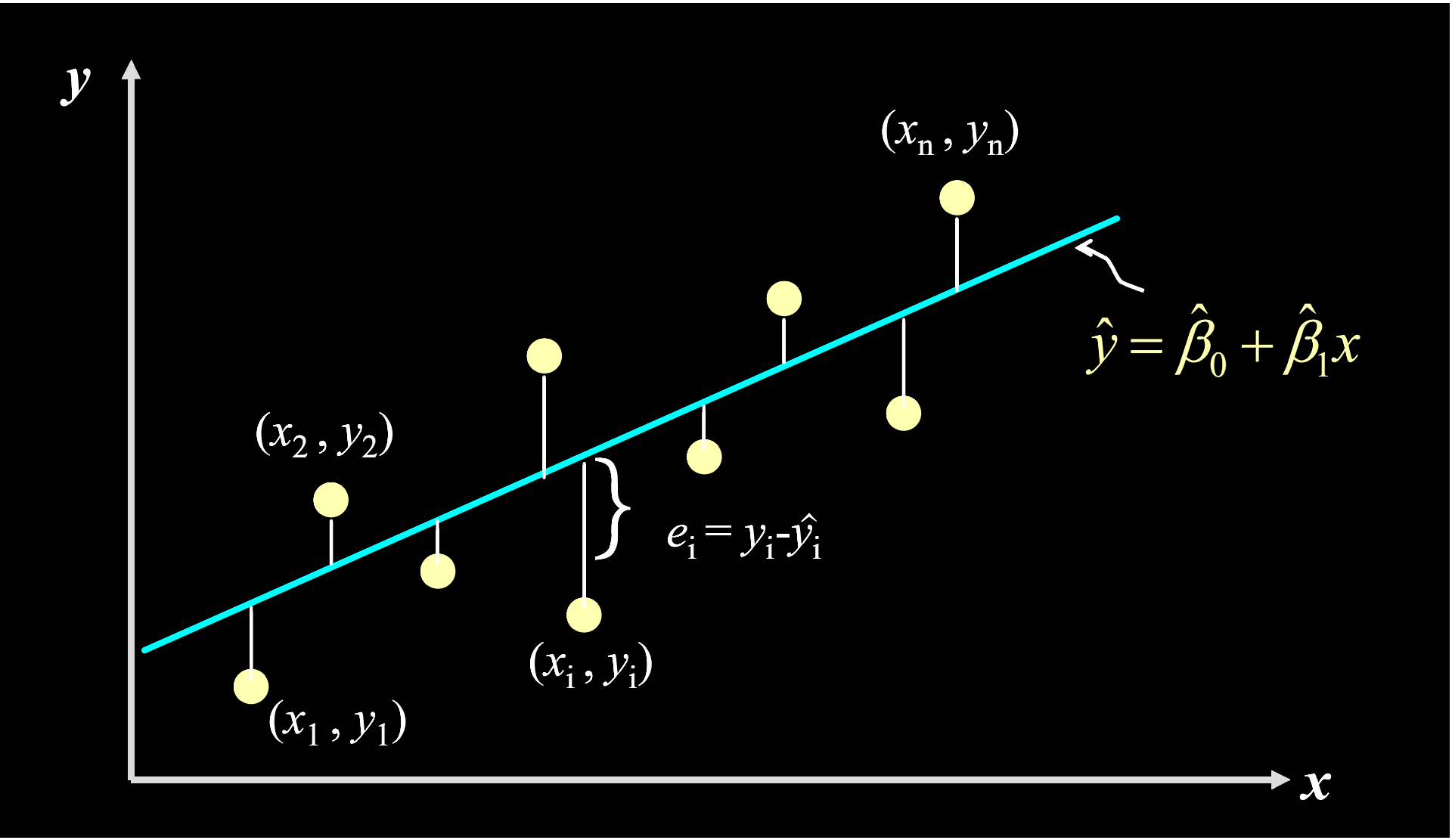
$$Q(\hat{\beta}_0,\hat{\beta}_1)=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$$

$$=\min_{\beta_0,\beta_1}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$ 称为y_i的回归拟合值,简称回归值或拟合值

$e_i=y_i-\hat{y}_i$ 称为y_i的残差

2.2 参数β_0、β_1的估计



2.2 参数β_0、β_1的估计

$$\begin{cases} \left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0 \\ \left.\dfrac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)x_i=0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases}$$

经整理后,得正规方程组

$$\begin{cases} n\hat{\beta}_0 + (\sum_{i=1}^{n} x_i)\hat{\beta}_1 = \sum_{i=1}^{n} y_i \\ (\sum_{i=1}^{n} x_i)\hat{\beta}_0 + (\sum_{i=1}^{n} x_i^2)\hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

2 .2 参数β_0、β_1的估计

得OLSE 为

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases}$$

记

$$L_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

$$L_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = L_{xy} / L_{xx} \end{cases}$$

2 .2 参数β_0、β_1的估计

续例2.1 $\bar{x}=\frac{49.2}{15}=3.28$, $\bar{y}=\frac{396.2}{15}=26.413$

$$L_{xx}=\sum_{i=1}^{n}x_i^2-n(\bar{x})^2$$

$$=196.16-15(3.28)^2=34.784$$

$$L_{xy}=\sum_{i=1}^{n}x_i y_i-n\bar{x}\,\bar{y}$$

$$=1470.65-1299.536=171.114$$

$$\begin{cases}\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=26.413-4.919\times3.28=10.279\\ \hat{\beta}_1=L_{xy}/L_{xx}=171.114/34.784=4.919\end{cases}$$

$$\hat{y}=10.279+4.919x$$

回归方程

2 .2 参数β_0、β_1的估计

二、最大似然估计

连续型：是样本的联合密度函数：
离散型：是样本的联合概率函数。
似然函数并不局限于独立同分布的样本。

似然函数

在假设$\varepsilon_i \sim N(0,\sigma^2)$时,由（2.10）式知$y_i$服从如下正态分布:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

2.2 参数β_0、β_1的估计

二、最大似然估计

$y_1, y_2, \ldots, y_n$
的似然函数为：

$$L(\beta_0,\beta_1,\sigma^2)=\prod_{i=1}^{n} f_i(y_i)$$

$$=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2\}$$

$$\ln(L)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2$$

对数似然
函数为：

与最小二乘原理完全相同

2.3 最小二乘估计的性质

一、线性

$\hat{\beta}_0$、$\hat{\beta}_1$是$y_1, y_2, \ldots, y_n$的线性函数：

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

2.3 最小二乘估计的性质

二、无偏性

$$E(\hat{\beta}_1)=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}E(y_i)$$

$$=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}(\beta_0+\beta_1 x_i)$$

$$=\beta_1$$

其中用到

$$\sum(x_i-\bar{x})=0$$

$$\sum(x_i-\bar{x})x_i=\sum(x_i-\bar{x})^2$$

2.3 最小二乘估计的性质

三、$\hat{\beta}_0$、$\hat{\beta}_1$ 的方差

$$\mathrm{var}(\hat{\beta}_1)=\sum_{i=1}^{n}\left[\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right]^2\mathrm{var}(y_i)=\frac{\sigma^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}$$

$$\mathrm{var}(\hat{\beta}_0)=\left[\frac{1}{n}+\frac{(\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]\sigma^2$$

$$\mathrm{cov}(\hat{\beta}_0,\hat{\beta}_1)=-\frac{\bar{x}}{L_{xx}}\sigma^2$$

2.3 最小二乘估计的性质

三、$\hat{\beta}_0$、$\hat{\beta}_1$ 的方差

$$\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{(\bar{x})^2}{L_{xx}})\sigma^2)$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$$

在正态假设下

Gauss　Markov条件

$$\begin{cases} E(\varepsilon_i) = 0, \quad i = 1, 2, \cdots, n \\ \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} \end{cases} \qquad (i, j = 1, 2, \cdots, n)$$

2.4 回归方程的显著性检验

一、*t* 检验

原假设： H_0：$\beta_1=0$

对立假设： H_1：$\beta_1\neq 0$

由 $$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$$

当原假设H_0：$\beta_1=0$成立时有：

$$\hat{\beta}_1 \sim N(0, \frac{\sigma^2}{L_{xx}})$$

2.4 回归方程的显著性检验

一、*t* 检验

构造*t* 统计量

$$t=\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/L_{xx}}}=\frac{\hat{\beta}_1\sqrt{Lxx}}{\hat{\sigma}}$$

其中 $$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}e_i^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

2.4 回归方程的显著性检验

二、用统计软件计算

1．例2.1 用Excel软件计算

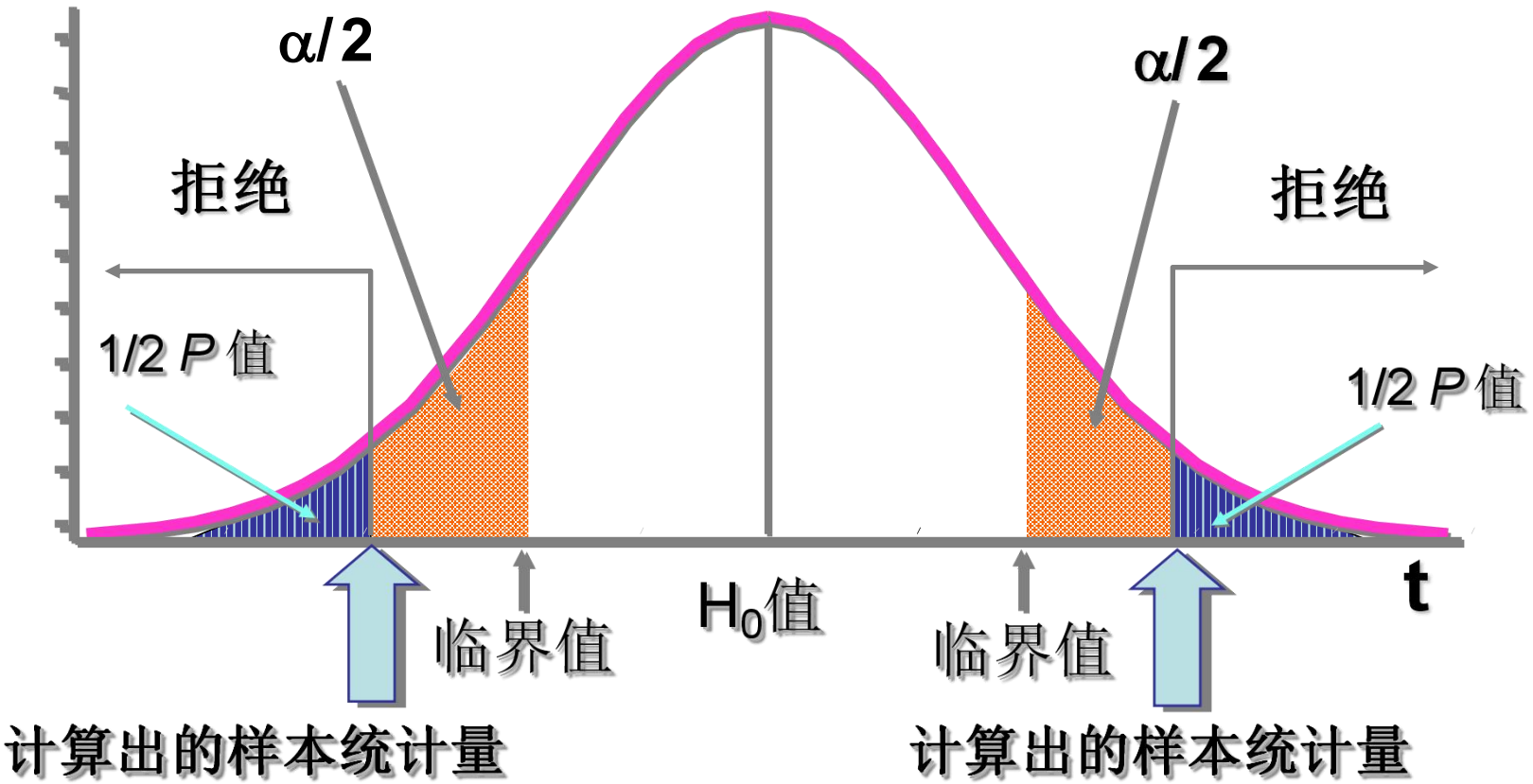
SUMMARY OUTPUT						
回归统计						
Multiple R	0.960977715					
R Square	0.923478169					
Adjusted R Square	0.917591874					
标准误差	2.316346184					
观测值	15					
方差分析						
	df	SS	MS	F	Significance F	
回归分析	1	841.766358	841.766358	156.8861596	1.2478E-08	
残差	13	69.75097535	5.365459643			
总计	14	911.5173333				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	10.27792855	1.420277811	7.236562082	6.58556E-06	7.20960489	13.34625221
x	4.919330727	0.392747749	12.52542054	1.2478E-08	4.070850801	5.767810653

什么是*P* 值?
(*P*-value)

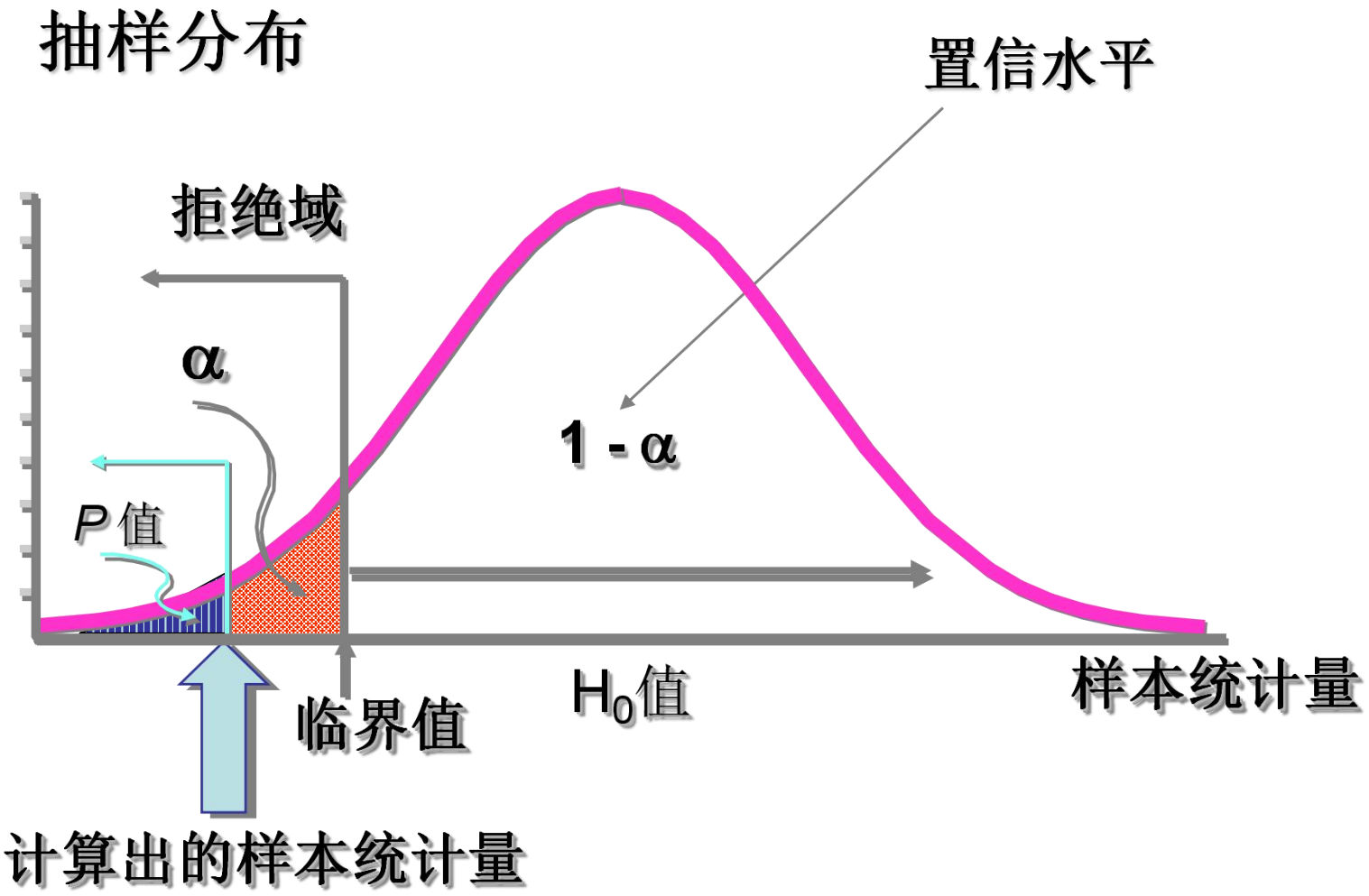
- *P* 值即显著性概率值

  Significence Probability Value
- 是当原假设为真时得到比目前的 样本更极端的样本的 概率，所谓极端就是与原假设相背离
- 它是用此样本拒绝原假设所犯弃真错误的真实概率，被称为观察到的(或实测的)显著性水平

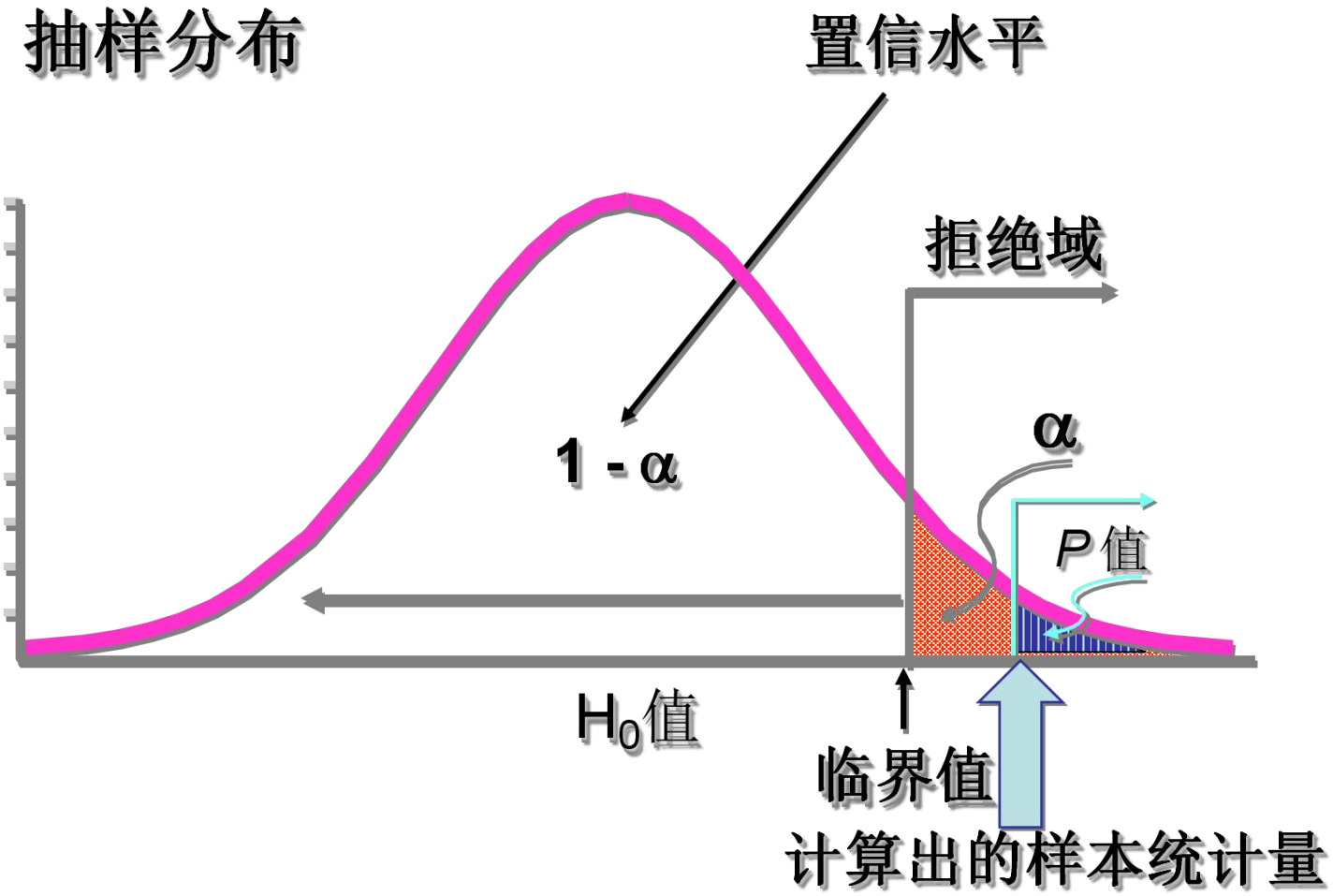
双侧检验的*P* 值



左侧检验的*P* 值



右侧检验的*P* 值



利用 *P* 值进行检验的决策准则

若p-值 ≥ α, 不能拒绝 H_0

若p-值 ＜ α, 拒绝 H_0

双侧检验p-值 =2×单侧检验p-值

2.4 回归方程的显著性检验

二、用统计软件计算

2. 例2.1用SPSS软件计算

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	x[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.961[a]	.923	.918	2.31635

a. Predictors: (Constant), x

2.4 回归方程的显著性检验

二、用统计软件计算

2.用SPSS软件计算

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	841.766	1	841.766	156.886	.000[a]
	Residual	69.751	13	5.365		
	Total	911.517	14			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.278	1.420		7.237	.000
	X	4.919	.393	.961	12.525	.000

a. Dependent Variable: Y

2.4 回归方程的显著性检验

三、F检验

平方和分解式

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

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

构造F检验统计量

$$F=\frac{SSR/1}{SSE/(n-2)}$$

2.4 回归方程的显著性检验

三、F检验

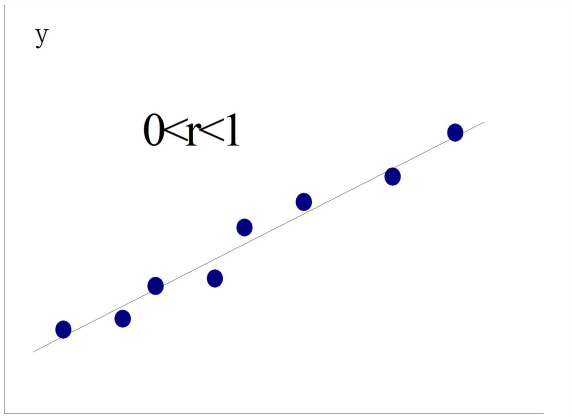
一元线性回归方差分析表

方差来源	自由度	平方和	均方	F值	P值
回归	1	SSR	$SSR/1$	$\dfrac{SSR/1}{SSE/(n-2)}$	$P(F>\text{F值})$ =P值
残差	n-2	SSE	$SSE/$（n-2）		
总和	n-1	SST			

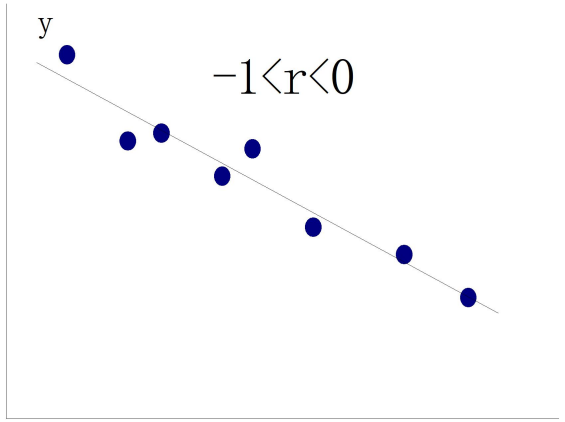
2.4 回归方程的显著性检验

四、相关系数的显著性检验

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\hat{\beta}_1\sqrt{\frac{L_{xx}}{L_{yy}}}$$



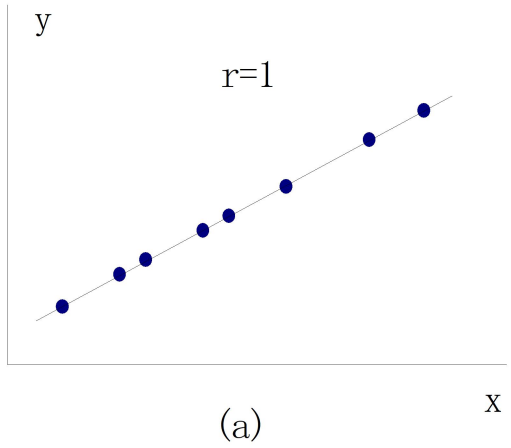
(e)



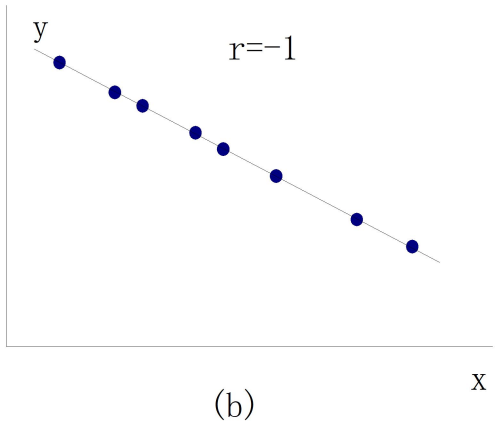
(f)

2.4 回归方程的显著性检验

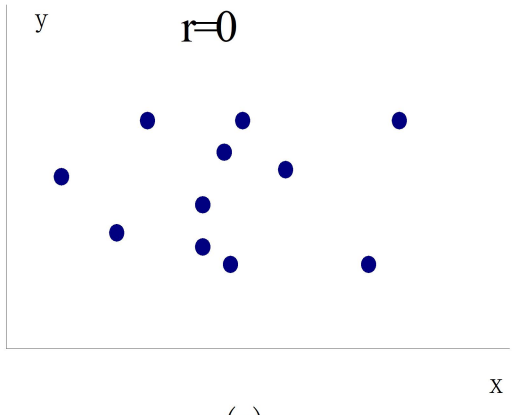
四、相关系数的显著性检验



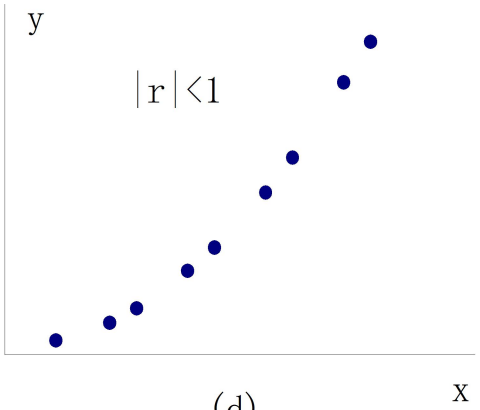
(a)



(b)



(c)



(d)

2.4 回归方程的显著性检验

四、相关系数的显著性检验

附表1　相关系数ρ=0的临界值表

n-2	5%	1%	n-2	5%	1%	n-2	5%	1%
1	0.997	1.000	16	0.468	0.590	35	0.325	0.418
2	0.950	0.990	17	0.456	0.575	40	0.304	0.393
3	0.878	0.959	18	0.444	0.561	45	0.288	0.372
4	0.811	0.947	19	0.433	0.549	50	0.273	0.354
5	0.754	0.874	20	0.423	0.537	60	0.250	0.325
6	0.707	0.834	21	0.413	0.526	70	0.232	0.302
7	0.666	0.798	22	0.404	0.515	80	0.217	0.283
8	0.632	0.765	23	0.396	0.505	90	0.205	0.267
9	0.602	0.735	24	0.388	0.496	100	0.195	0.254
10	0.576	0.708	25	0.381	0.487	125	0.174	0.228
11	0.553	0.684	26	0.374	0.478	150	0.159	0.208
12	0.532	0.661	27	0.367	0.470	200	0.138	0.181
13	0.514	0.641	28	0.361	0.463	300	0.113	0.148
14	0.497	0.623	29	0.355	0.456	400	0.098	0.128
15	0.482	0.606	30	0.349	0.449	1000	0.062	0.081

2.4 回归方程的显著性检验

四、相关系数的显著性检验

$$t = \frac{\sqrt{n-2}\ r}{\sqrt{1-r^2}}$$

用**SPSS**软件做相关系数的显著性检验

Correlations

		Y	X
Y	Pearson Correlation	1.000	.961
	Sig. (2-tailed)	.	.000
	N	15	15
X	Pearson Correlation	.961	1.000
	Sig. (2-tailed)	.000	.
	N	15	15

2.4 回归方程的显著性检验

四、相关系数的显著性检验

两变量间相关程度的强弱分为以下几个等级：

当$|r|\geq 0.8$时，视为高度相关；

当$0.5\leq|r|< 0.8$时，视为中度相关；

当$0.3\leq|r|< 0.5$时，视为低度相关；

当$|r|< 0.3$时，表明两个变量之间的相关程度极弱，在实际应用中可视为不相关。

2.4 回归方程的显著性检验

五、三种检验的关系

H_0: β=0

$$t=\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/L_{xx}}}=\frac{\hat{\beta}_1\sqrt{Lxx}}{\hat{\sigma}}$$

H_0: ρ=0

$$t=\frac{\sqrt{n-2}\ r}{\sqrt{1-r^2}}$$

H_0: 回归无效

$$F=\frac{SSR/1}{SSE/(n-2)}$$

2.4 回归方程的显著性检验

六、样本决定系数

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

可以证明 $$r^2 = \frac{SSR}{SST} = \frac{L_{xy}^2}{L_{xx}\,L_{yy}} = (r)^2$$

2.5 残差分析

一、残差概念与残差图

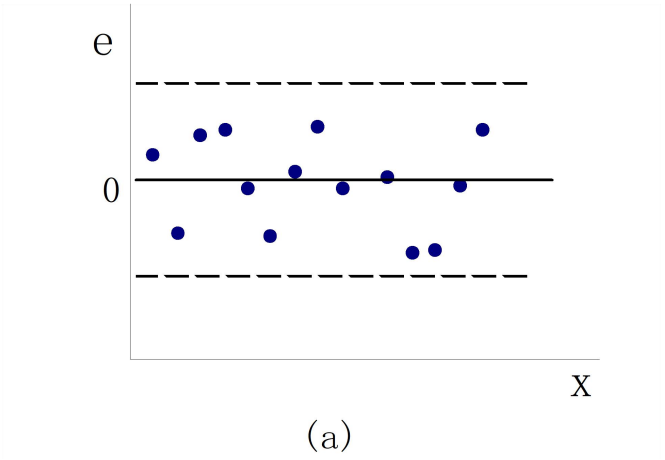
残差 $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

误差项 $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$

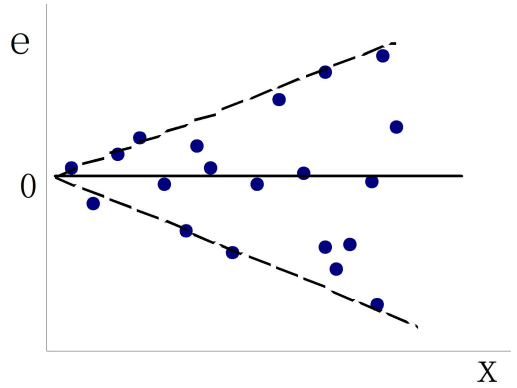
残差e_i是误差项ε_i的估计值。

2.5 残差分析

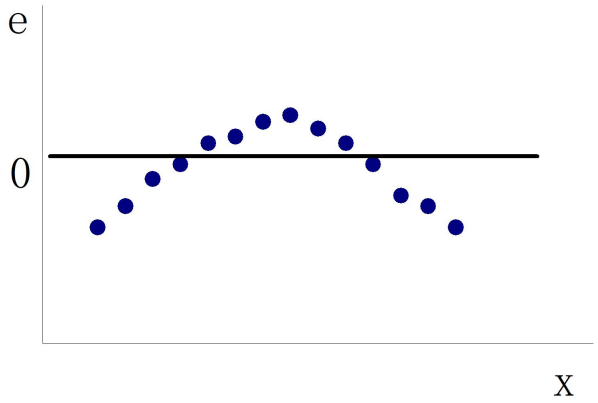
一、残差概念与残差图



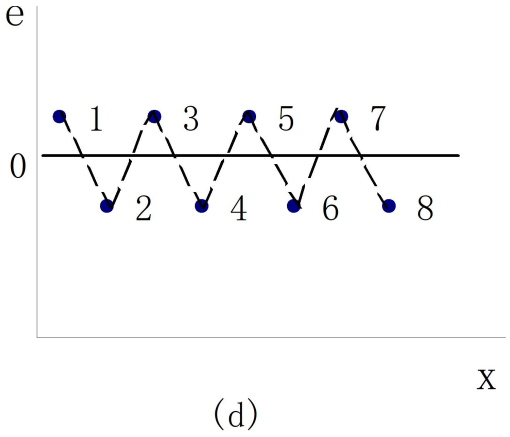
(a)



(b)



(c)



(d)

2.5 残差分析

一、残差概念与残差图

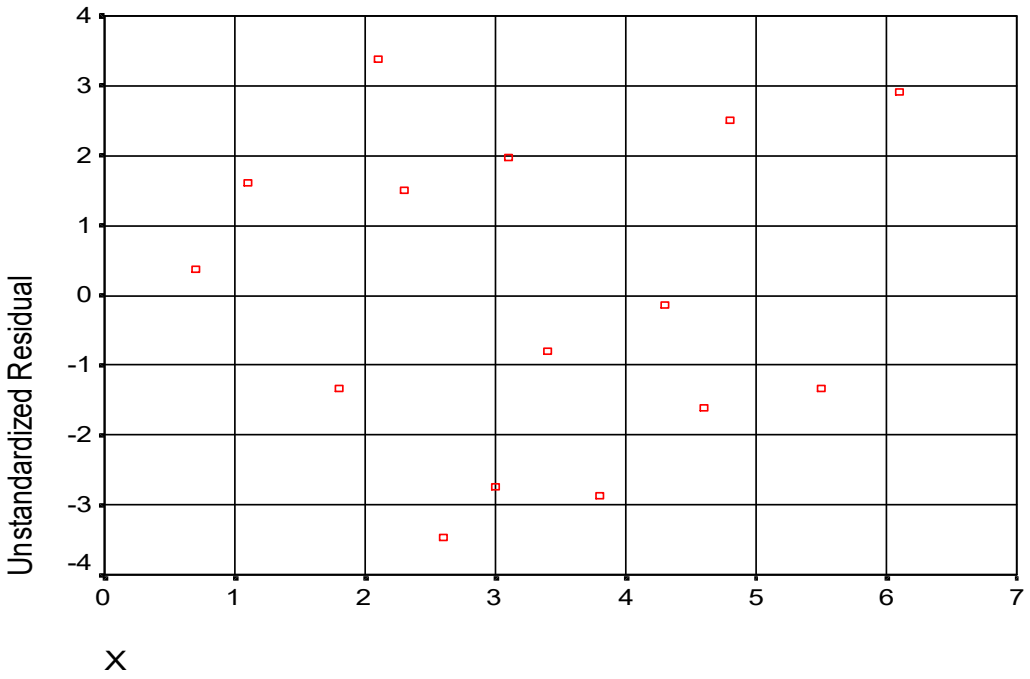


图 2.6 火灾损失数据残差图

2.5 残差分析

二、残差的性质

性质1　$\mathrm{E}(e_i)=0$

证明：
$$E(e_i)=E(y_i)-E(\hat{y}_i)$$
$$=(\beta_0+\beta_1 x_i)-E(\hat{\beta}_0+\hat{\beta}_1 x_i)$$
$$=0$$

2.5 残差分析

二、残差的性质

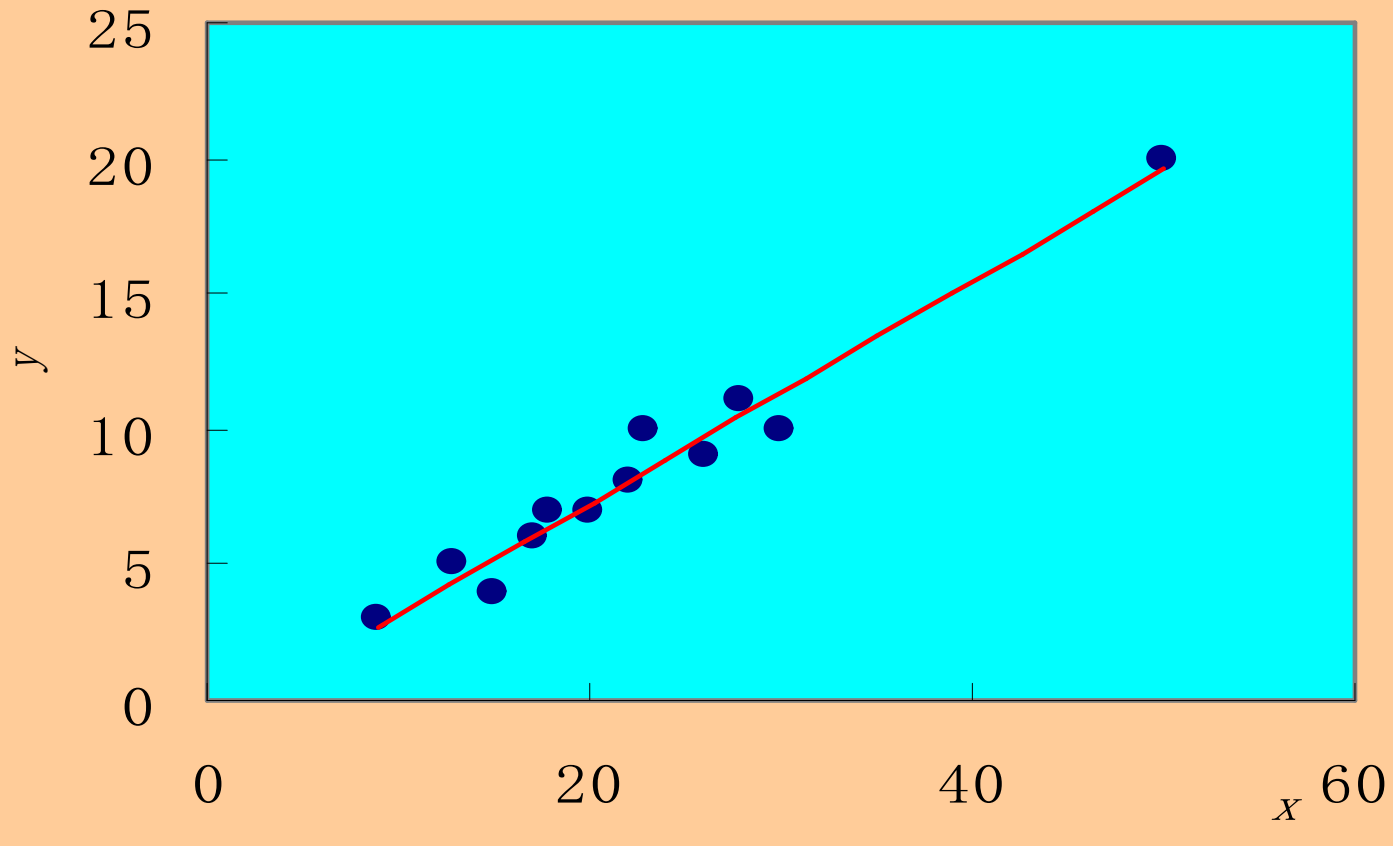
性质2 $$\operatorname{var}(e_i)=\left[1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{L_{xx}}\right]\sigma^2$$

$$=(1-h_{ii})\,\sigma^2$$

其中 $h_{ii}=\dfrac{1}{n}+\dfrac{(x_i-\bar{x})^2}{L_{xx}}$ 称为杠杆值

2.5 残差分析

二、残差的性质



存在高杠杆率观测值的散点图

2.5 残差分析

二、残差的性质

性质3. 残差满足约束条件:

$$\sum_{i=1}^{n} e_i = 0$$

$$\sum_{i=1}^{n} x_i e_i = 0$$

2.5 残差分析

三、改进的残差

标准化残差 $ZRE_i = \dfrac{e_i}{\hat{\sigma}}$

学生化残差 $SRE_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

2.6 回归系数的区间估计

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$$

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

$$P\left(\left|\frac{(\hat{\beta}_1 - \beta_1)\sqrt{L_{xx}}}{\hat{\sigma}}\right| < t_{\alpha/2}(n-2)\right) = 1-\alpha$$

等价于 $P(\hat{\beta}_1 - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{L_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{L_{xx}}}) = 1-\alpha$

$$(\hat{\beta}_1 - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{L_{xx}}}, \hat{\beta}_1 + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{L_{xx}}})$$

β_1的1-α置信区间

2.7 预测和控制

一、单值预测

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$E(\hat{y}_0) = E(y_0) = \beta_0 + \beta_1 x_0$$

2.7 预测和控制

二、区间预测

1．因变量新值的区间预测

找一个区间（T_1, T_2），使得

$$P(T_1 < y_0 < T_2) = 1 - \alpha$$

需要首先求出其估计值 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 的分布

二、区间预测　1 因变量新值的区间预测

以下计算 $\hat{y}_0$ 的方差

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}}\right) y_i$$

$$\mathrm{var}(\hat{y}_0) = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}}\right)^2 \mathrm{var}(y_i) = \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}\right) \sigma^2$$

从而得　　$\hat{y}_0 \sim N\left(\beta_0 + \beta_1 x_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}\right)\sigma^2\right)$

二、区间预测　1 因变量新值的区间预测

记　$h_{00}=\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{L_{xx}}$

则　$\hat{y}_0 \sim N(\beta_0+\beta_1 x_0, h_{00}\sigma^2)$

$$\mathrm{var}(y_0-\hat{y}_0)=\mathrm{var}(y_0)+\mathrm{var}(\hat{y}_0)=\sigma^2+h_{00}\sigma^2$$

于是有　$y_0-\hat{y}_0 \sim N(0,(1+h_{00})\sigma^2)$

$$t=\frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\ \hat{\sigma}} \sim t(n-2)$$

二、区间预测　1 因变量新值的区间预测

$$P\left(\left|\frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\,\hat{\sigma}}\right|\le t_{\alpha/2}(n-2)\right)=1-\alpha$$

y_0的置信概率为1-α的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{1+h_{00}}\,\hat{\sigma}$$

y_0的置信度为95%的置信区间近似为

$$\hat{y}_0 \pm 2\,\hat{\sigma}$$

二、区间预测 2 因变量平均值的区间估计

$E(y_0)=\beta_0+\beta_1 x_0$是常数

$$\hat{y}_0 - E(y_0) \sim N(0,(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}})\sigma^2)$$

得$E(y_0)$的1-α的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{h_{00}}\ \hat{\sigma}$$

二、区间预测 计算

对例2.1的火灾损失数据，假设保险公司希望预测一个距最近的消防队x_0=3.5公里的居民住宅失火的损失

点估计值 $\hat{y}_0 = 10.278 + 4.919 \times 3.5 = 27.50$

95%区间估计 单个新值： （22.32，32.67）

平均值E(y_0)： （26.19，28.80）

$\hat{y}_0$ 的95%的近似置信区间为

$$(\hat{y}_0 - 2\hat{\sigma},\quad \hat{y}_0 + 2\hat{\sigma})$$

=（27.50-2×2.316，27.50+2×2.316）

=（22.87，32.13）

三、控制问题

给定y的预期范围(T_1, T_2),如何控制自变量x的值才能以1-α的概率保证

$$P(T_1 < y < T_2) = 1 - \alpha$$

用近似的预测区间来确定x。如果α=0.05,则要求

$$\begin{cases} \hat{y}(x) - 2\hat{\sigma} > T_1 \\ \hat{y}(x) + 2\hat{\sigma} < T_2 \end{cases}$$

把 $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ 带入

当$\hat{\beta}_1 > 0$时，得 $\dfrac{T_1 + 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \dfrac{T_2 - 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1}$

当$\hat{\beta}_1 < 0$时，得 $\dfrac{T_2 - 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \dfrac{T_1 + 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1}$

2.8 本章小结与评注

一、一元线性回归模型从建模到应用的全过程

例2.2 全国人均消费金额记作y(元); 人均国民收入记为x(元)

表2.2　　人均国民收入表

年份	人均国民收入（元）	人均消费金额（元）	年份	人均国民收入（元）	人均消费金额（元）
1980	460	234.75	1990	1634	797.08
1981	489	259.26	1991	1879	890.66
1982	525	280.58	1992	2287	1063.39
1983	580	305.97	1993	2939	1323.22
1984	692	347.15	1994	3923	1736.32
1985	853	433.53	1995	4854	2224.59
1986	956	481.36	1996	5576	2627.06
1987	1104	545.40	1997	6053	2819.36
1988	1355	687.51	1998	6392	2958.18
1989	1512	756.27			

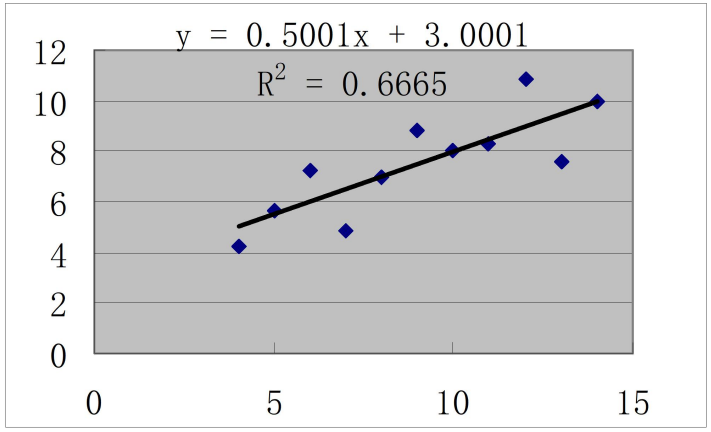
2.8 本章小结与评注

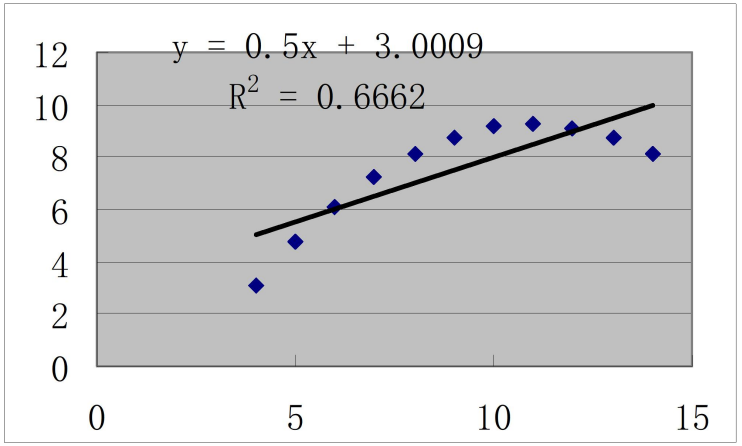
二、有关回归假设检验问题

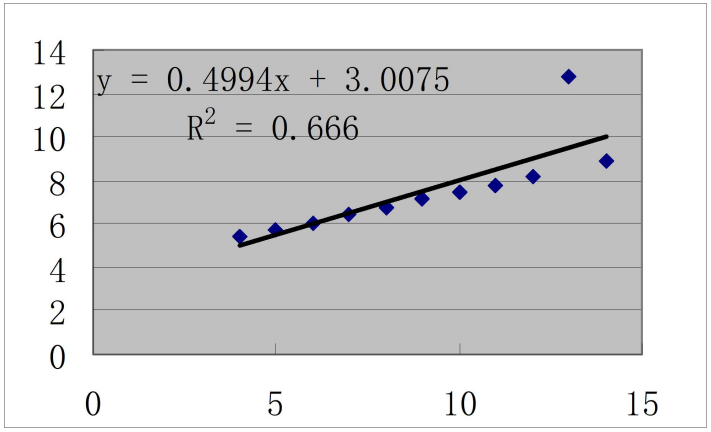
1973年Anscombe构造了四组数据, 这四组数据所建的回归方程是相同的,决定系数,F统计量也都相同,且均通过显著性检验。

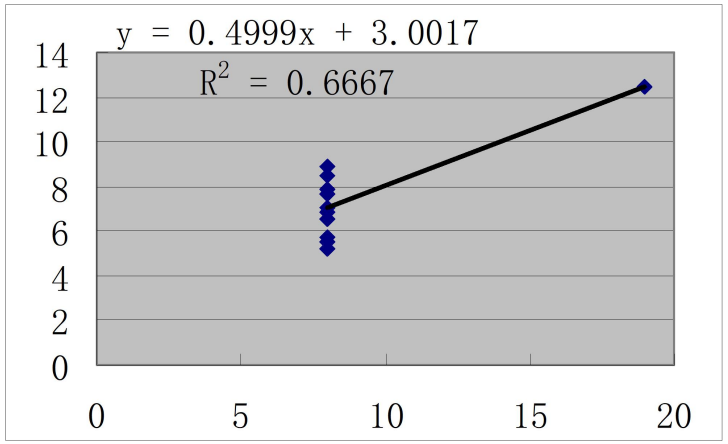
第一组		第二组		第三组		第四组	
x	y	x	y	x	y	x	y
4	4.26	4	3.1	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.44	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.1	14	8.84	19	12.5

2.8 本章小结与评注


y = 0.5001x + 3.0001
R2 = 0.6665
12
10
8
6
4
2
0
0
5
10
15


y = 0.5x + 3.0009
R2 = 0.6662
12
10
8
6
4
2
0
0
5
10
15


y = 0.4994x + 3.0075
R2 = 0.666
14
12
10
8
6
4
2
0
0
5
10
15


y = 0.4999x + 3.0017
R2 = 0.6667
14
12
10
8
6
4
2
0
0
5
10
15
20