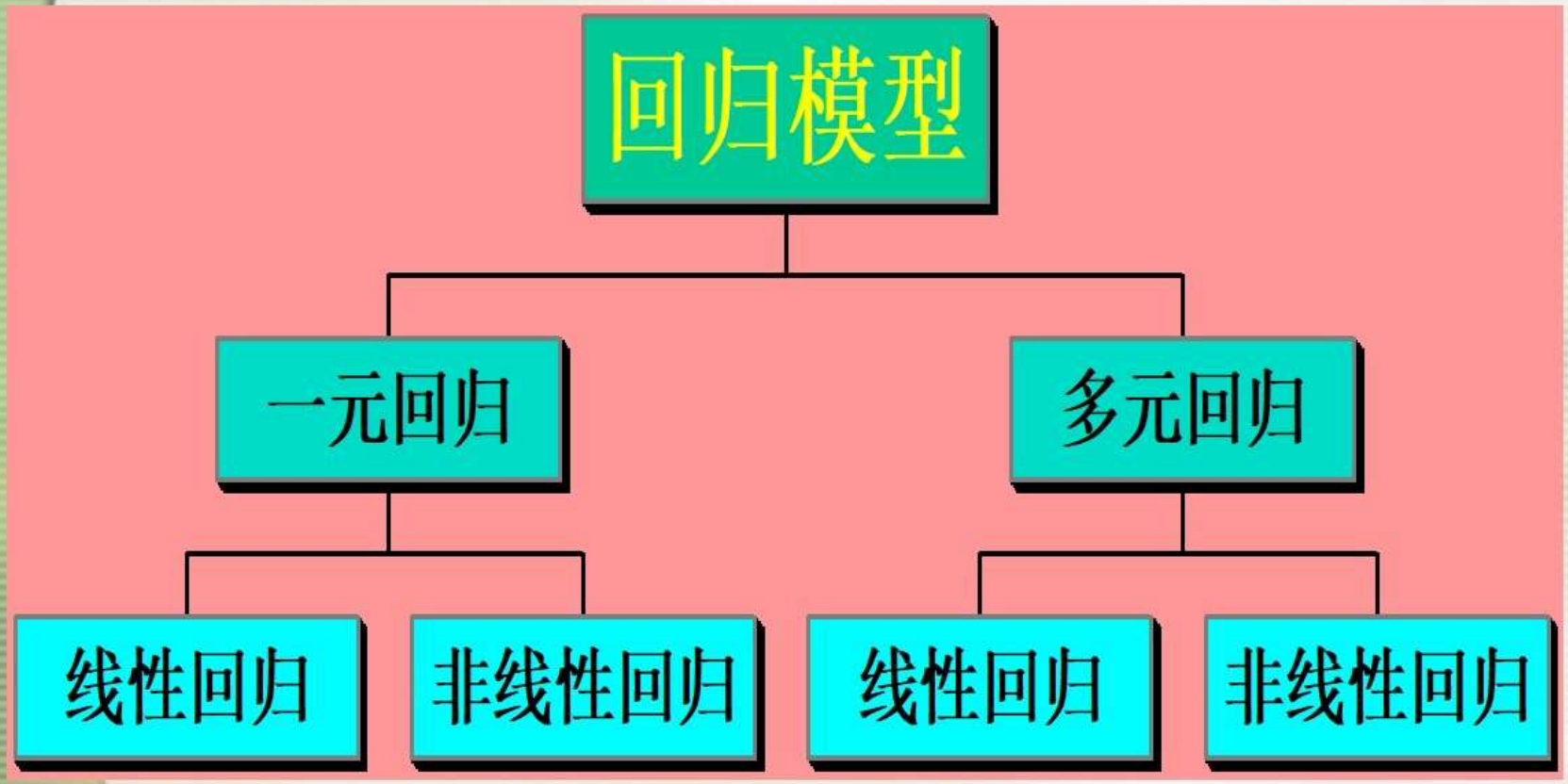
第三节　一元线性回归

- 一、一元线性回归
- 二、a, b的估计
- 三、总体方差的估计
- 四、线性假设的显著性检验
- 五、系数b的置信区间
- 六、回归预测
- 七、可化为一元线性回归的例子(自学)

•回归模型的类型



•一、一元线性回归

1. 只涉及一个自变量的回归；
2. 因变量y与自变量x之间为线性关系。
   - 被预测或被解释的变量称为因变量**(dependent variable)**，用y表示；
   - 用来预测或用来解释因变量的一个或多个变量称为自变量**(independent variable)**，用x表示。
3. 因变量与自变量之间的关系用一个线性方程来表示。

一元线性回归模型的基本形式

①描述因变量 y 如何依赖于自变量 x 和误差项 ε 的方程称为理论回归模型

②一元线性回归模型可表示为

$y_i = a + bx_i + \varepsilon_i, \varepsilon_i \square N(0, \sigma^2)$, 各$\varepsilon_i$相互独立

理论回归模型

- y 是 x 的线性函数(部分)加上随机误差项
- 线性部分反映了由于 x 的变化而引起的 y 的变化；误差项 ε 是随机变量(未纳入模型但对y有影响的诸多因素的综合影响)，反映了除 x 和 y 之间的线性关系之外的随机因素对 y 的影响，是不能由 x 和 y 之间的线性关系所解释的变异性。
- a和 b称为模型的参数

- 在抽样中，自变量x的取值是固定的，即x是非随机的；因变量y是随机的。

  即当解释变量X取某固定值时，Y的值不确定，Y的不同取值形成一定的分布，这是Y的条件分布。

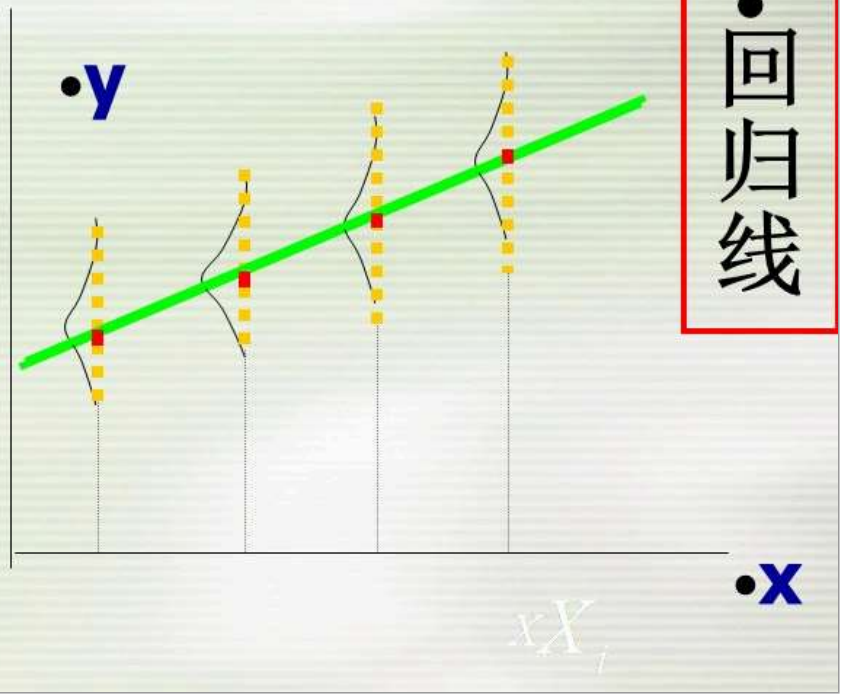
回归线，描述的是Y的条件期望E（Y/xi）与之对应xi，代表这些Y的条件期望的点的轨迹所形成的直线或曲线。

如

$$E(y_i) = a + bx_i$$

注意：

由于单个数据点是从y的分布中抽出来的，可能不在这条回归线上，因此必须包含随机误差项ε来描述模型数据点.



回归模型的基本假设

假设1：误差项的期望值为0，即对所有的i有 $E(\varepsilon_i)=0$

假设2：误差项的方差为常数，即对所有的i有 $\mathrm{var}(\varepsilon_i)=E(\varepsilon_i^2)=\sigma$

假设3：误差项之间不存在自相关关系，其协方差为0，

即当 $i\neq j$ 时，$\mathrm{cov}(\varepsilon_i,\varepsilon_j)=0$

有 ；

假设4：自变量是给定的变量，与随机误差项线性无关；

假设5：随机误差项服从正态分布。即 $\varepsilon \sim N(0,\sigma^2)$

以上这些基本假设是德国数学家高斯最早提出的，故也称为高斯假定或标准假定。

•回归方程(regression equation)

1. 描述 y 的平均值或期望值如何依赖于 x 的方程称为回归方程
2. 一元线性回归方程的形式如下:

$$\hat{y} = a + bx$$

- 方程的图示是一条直线，也称为直线回归方程。
- **a**是回归直线在 y 轴上的截距，是当 x=0 时 y 的期望值;
- ***b***是直线的斜率，称为回归系数，表示当 x 每变动一个单位时，y的平均变动值。

•.估计的回归方程(estimated regression equation)

1. 总体回归参数 a 和 b 是未知的，必须利用样本数据去估计；

2. 用样本统计量 $\hat{a}$，$\hat{b}$ 代替回归方程中的未知参数 a 和 b，就得到了估计的回归方程.

3. 一元线性回归中估计的回归方程为

$$\hat{y} = \hat{a} + \hat{b}x$$

•其中：$\hat{a}$ 是估计的回归直线在 y 轴上的截距，$\hat{b}$ 是直线的斜率，它表示对于一个给定的 x 的值，$\hat{y}$ 是 y 的估计值，也表示 x 每变动一个单位时，y 的平均变动值 。

二、a, b的估计（普通最小二乘估计法）
(ordinary least squares estimato

1. 使因变量的观察值与估计值之间的离差平方和达到最小来求得 $\hat{a}$ 和 $\hat{b}$ 的方法。即

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 = \text{最小值}$$

2. 用最小平方法拟合的直线来代表x与y之间的关系与实际数据的误差比其他任何直线的误差都小。

参数的最小二乘估计

$$\hat{b} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

2025/3/17

242

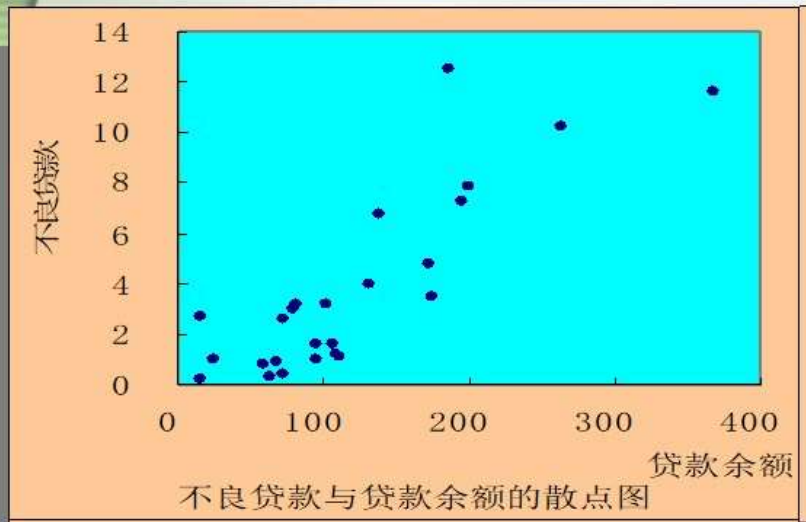
例9.6

【例10.7】一家大型商业银行在多个地区设有分行，其业务主要是进行基础设施建设、国家重点项目建设、固定资产投资等项目的贷款。近年来，该银行的贷款额平稳增长，但不良贷款额也有较大比例的增长，这给银行业务的发展带来较大压力。为弄清不良贷款形成的原因，管理者希望利用银行业务的有关数据进行定量分析，以便找出控制不良贷款的办法。下面是该银行所属的25家分行2002年的有关业务数据

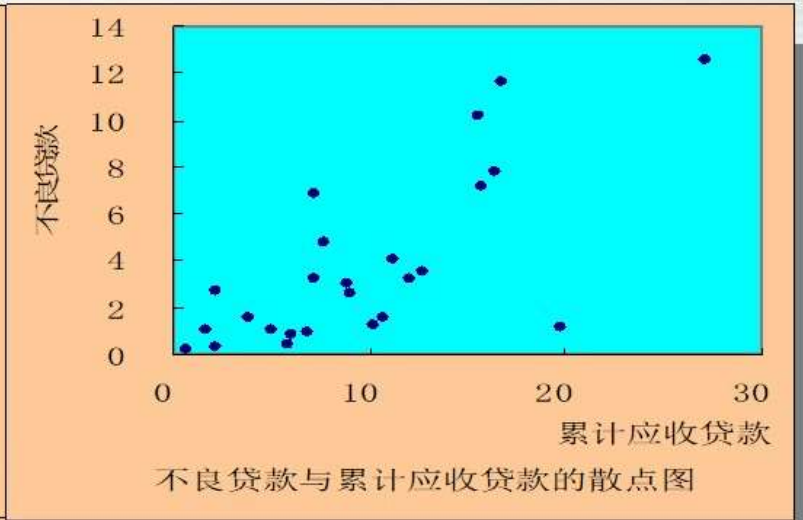
A	B	C	D	E	F
分行号	不良贷款（亿元）	各项贷款余额（亿元）	本年累计应收贷款（亿元）	贷款项目个数（个）	本年固定资产投资额（亿元）
1	0.9	67.3	6.8	8	51.9
2	10.1	111.3	19.8	16	90.9
3	4.8	173	7.7	17	73.7
4	3.2	80.8	7.2	10	14.5
5	17.8	199.7	16.5	19	63.2
6	2.7	16.2	2.2	1	2.2
7	1.6	107.4	10.7	17	20.2
8	12.5	185.4	27.1	18	43.8
9	7	96.1	1.7	10	55.9
10	2.6	72.8	9.1	14	64.3
11	0.3	64.2	2.1	11	42.7
12	14	132.2	11.2	23	76.7
13	0.8	58.6	6	14	22.8
14	13.5	174.6	12.7	26	117.1
15	30.2	263.5	15.6	34	146.7
16	8	79.3	8.9	15	29.9
17	5.2	14.8	0.3	2	42.1
18	8.4	73.5	5.9	11	25.3
19	10	24.7	5	4	13.4
20	6.8	139.4	7.2	28	64.3
21	31.6	368.2	16.8	32	163.9
22	1.6	95.7	3.8	10	44.5
23	9.2	109.6	10.3	14	67.9
24	17.2	196.2	15.8	16	39.7
25	8.2	102.2	12	10	97.1

2025/3/17 244

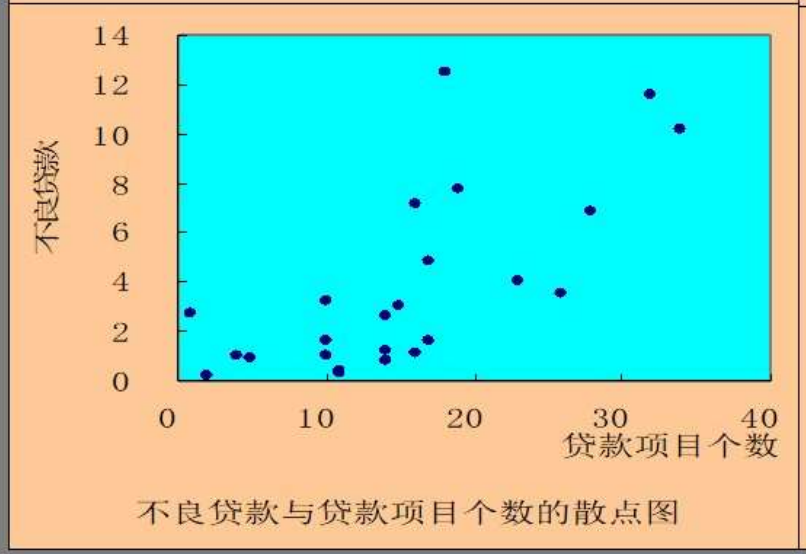
不良贷款对其他变量的散点图



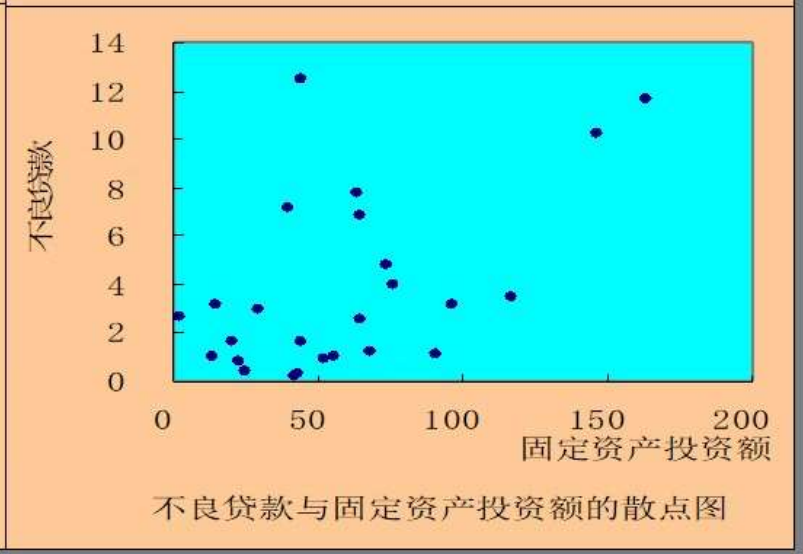
不良贷款与贷款余额的散点图



不良贷款与累计应收贷款的散点图



不良贷款与贷款项目个数的散点图



不良贷款与固定资产投资额的散点图

用Excel计算相关系数

	列 1	列 2	列 3	列 4	列 5
列 1	1				
列 2	0.849736	1			
列 3	0.613003	0.679407	1		
列 4	0.713496	0.851427	0.589245	1	
列 5	0.738537	0.779702	0.471902	0.755278	1

SUMMARY OUTPUT

回归统计

Multiple R	0.849736
R Square	0.722051
Adjusted R Square	0.709966
标准误差	4.45116
观测值	25

方差分析

	df	SS	MS	F	Significance F
回归分析	1	1183.795	1183.795	59.74896	7.69E-08
残差	23	455.6949	19.81282		
总计	24	1639.49			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-1.38473	1.625488	-0.85189	0.40306	-4.74731	1.977845	-4.74731	1.977845
X Variable 1	0.087411	0.011308	7.729745	7.69E-08	0.064018	0.110804	0.064018	0.110804

经验回归方程的求法

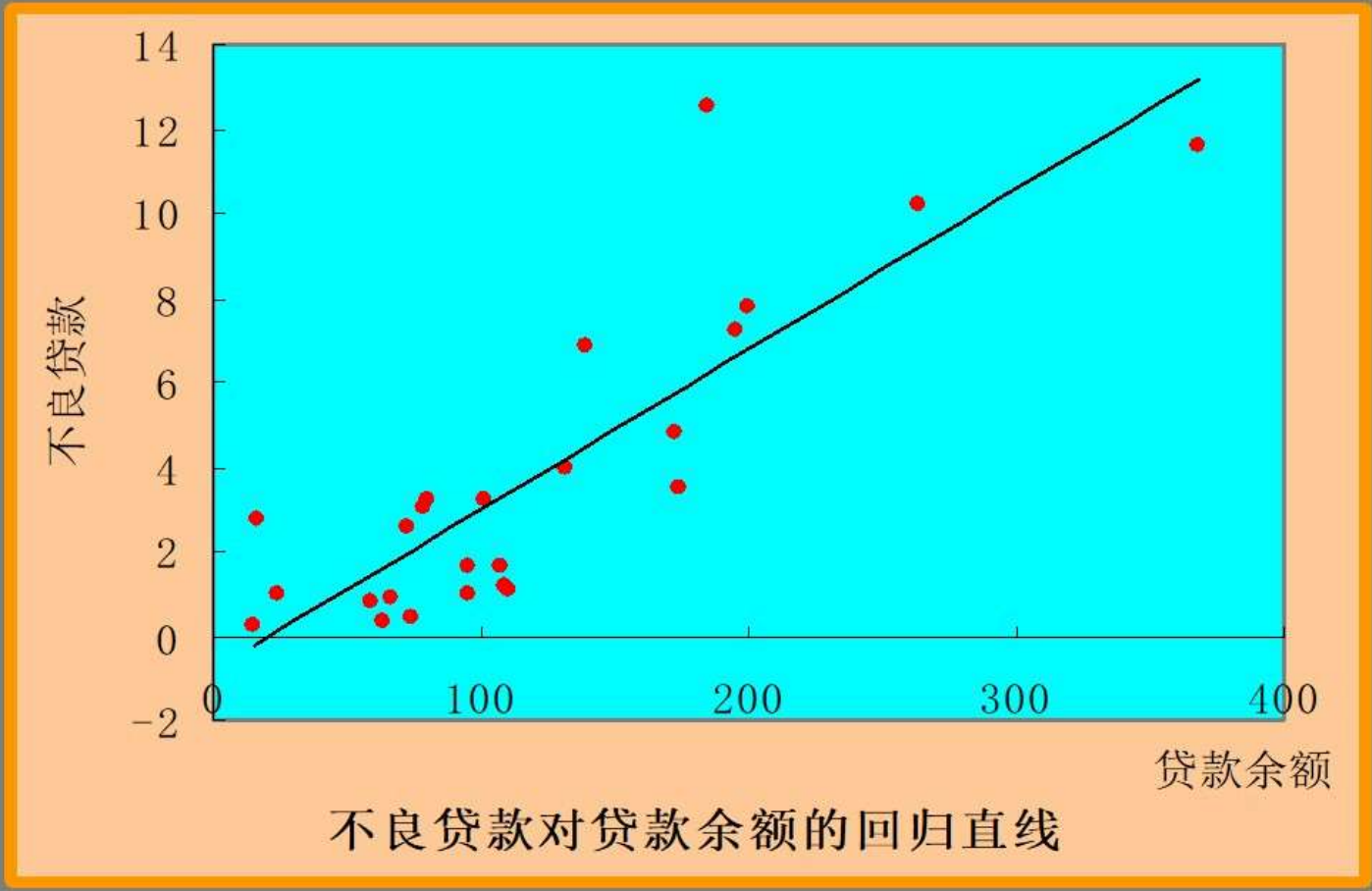
- 回归方程为：
- $\hat{y} = -1.38473 + 0.087411\,x$

- 回归系数 $\hat{b}$ =0.087411 表示，贷款余额每增加1亿元，不良贷款平均增加0.087411亿元

$\hat{\beta}_1$

估计回归方程的求法

不良贷款对贷款余额回归方程的图示



不良贷款对贷款余额的回归直线

三、σ^2 的估计

$$Q_e = \sum (y - \hat{y})^2 = \sum (y - \hat{a} - \hat{b}x)^2$$

称 为残差平方和，则

$$\hat{\sigma}^2 = \frac{Q_e}{n-2},$$

$$\hat{\sigma} = \sqrt{\frac{Q_e}{n-2}}$$

四、线性假设的显著性检验

$H_0: b = 0, H_1: b \neq 0$

$\hat{b} \sim N(b, \sigma^2 / S_{xx}), \dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2),$

$\dfrac{\hat{b}-b}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2)$。当$H_0$为真时，

$t = \dfrac{\hat{b}}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2)$, 拒绝域为$|t| \geq t_{\alpha/2}(n-2).$

五、系数b的置信区间P252

b的置信水平为$1-\alpha$的置信区间为：

$$\left(\hat{b} \pm t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)$$

六、回归函数$\mu(x)=a+bx$ 函数值的点估计和置信区间

- 回归函数的点估计值为

$$\hat{y}_0=\hat{\mu}(x_0)=\hat{a}+\hat{b}x$$

- $\mu(x_0)=a+bx$ 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\hat{a}+\hat{b}x_0\pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right)$$

七、Y的观测值的点预测和预测区间

$Y_0 = a + bx_0 + \varepsilon_0$的点预测值：$\hat{Y}_0 = \hat{a} + \hat{b}x_0$

Y_0的置信水平为$1-\alpha$的预测区间为

$$\left(\hat{a} + \hat{b}x_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

2025/3/17

254

第四节　多元线性回归

P257

- 因为客观现象非常复杂，现象之间的联系方式和性质各不相同，影响因变量变化的自变量往往是多个而不只是一个，其中既有主要因素也有次要因素。如果仅仅进行一元回归分析，不一定能得到满意的结果。因此，有必要将一个因变量与多个自变量联系起来进行分析。

多元线性回归

- 在线性相关条件下，研究两个和两个以上自变量对一个因变量的数量变化关系，称为多元线性回归分析，表现这一数量关系的数学表达式则称为多元线性回归方程或多元线性回归模型。

多元线性回归

多元线性回归模型:

$$Y = b_0 + b_1 x_1 + \cdots + b_p x_p + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

多元线性回归方程:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p$$

实验设计与分析