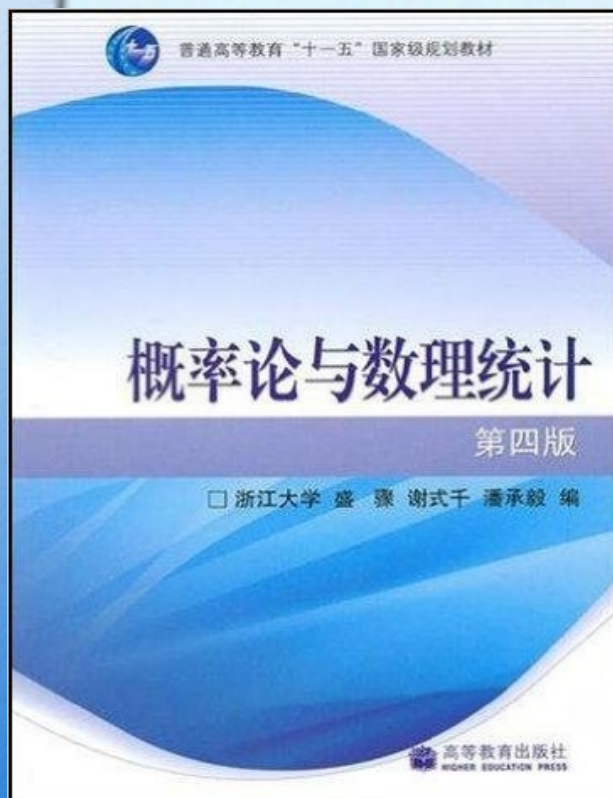


目 录



- 第一章 概率论的基本概念
- 第二章 随机变量及其分布
- 第三章 多维随机变量及其分布
- 第四章 随机变量的数字特征
- 第五章 大数定律及中心极限定理
- 第六章 样本及抽样分布
- 第七章 参数估计
- 第八章 假设检验
- 第九章 方差分析及回归分析

第九章 方差分析及回归分析

第一节 单因素试验的方差分析

第二节 双因素试验的方差分析

第三节 一元线性回归

第四节 多元线性回归

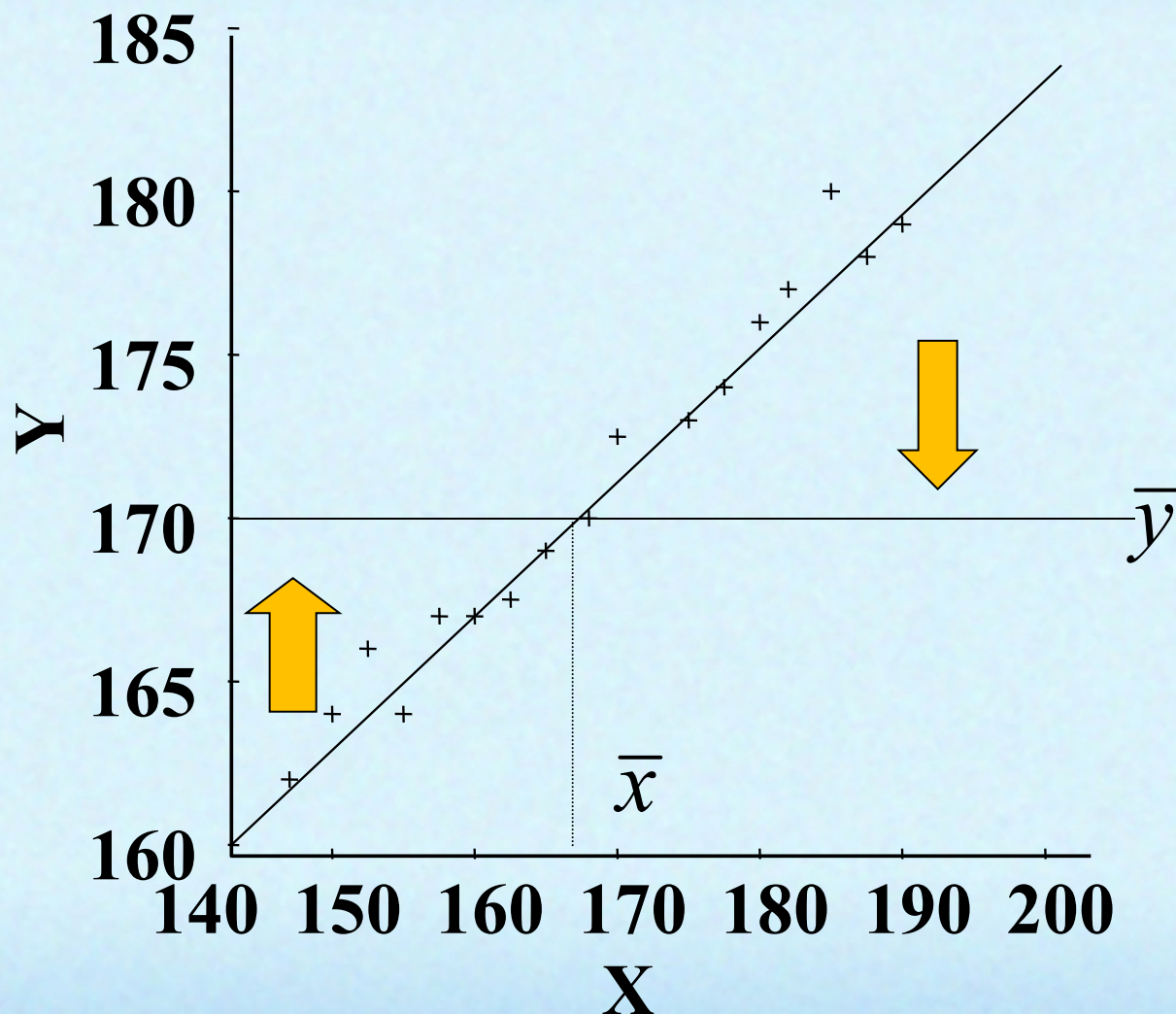
第三节 一元线性回归

- 一、回归分析的基本的思想
- 二、一元回归的数学模型
- 三、回归实验
- 四、小结与思考

父亲们的身高与儿子们的身高之间关系的研究

- 1889年高尔顿收集了上千个家庭的身高、臂长和腿长的记录
- 打算寻找出儿子们身高与父亲们身高之间关系的具体表现形式
- 下图是根据1078个家庭的调查所作的散点图（略图）

儿子们身高向着平均身高“回归”，以保持种族的稳定



“回归”一词的由来

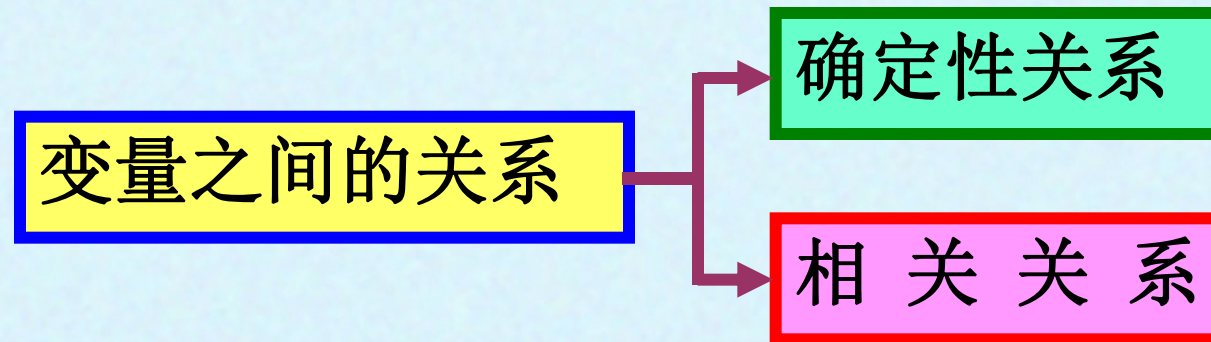
- 从图上虽可看出，个子高的父亲确有生出个子高的儿子的倾向，同样地，个子低的父亲确有生出个子低的儿子的倾向。得到的具体规律如下：

$$y = a + bx + \varepsilon$$

$$\hat{y} = 84.33 + 0.516x$$

- 发现某人种的平均身高是相当稳定的。最后得到结论：儿子们的身高回复于全体男子的平均身高，即“回归”

一、回归分析的基本思想



$$S = \pi r^2$$

确定性关系

身高和体重

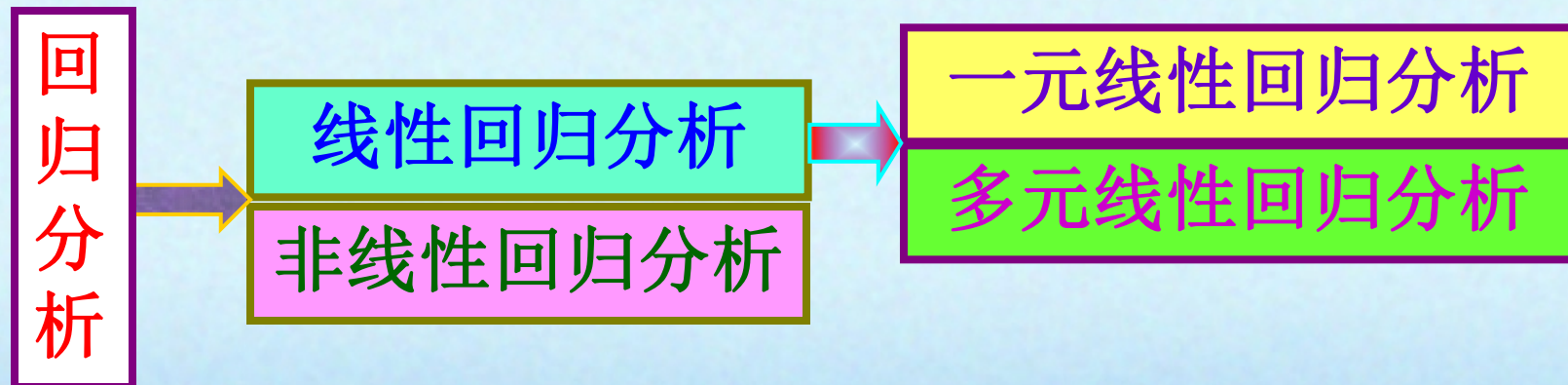
相关关系

相关关系的特征是：

变量之间的关系很难用一种精确的方法表示出来。

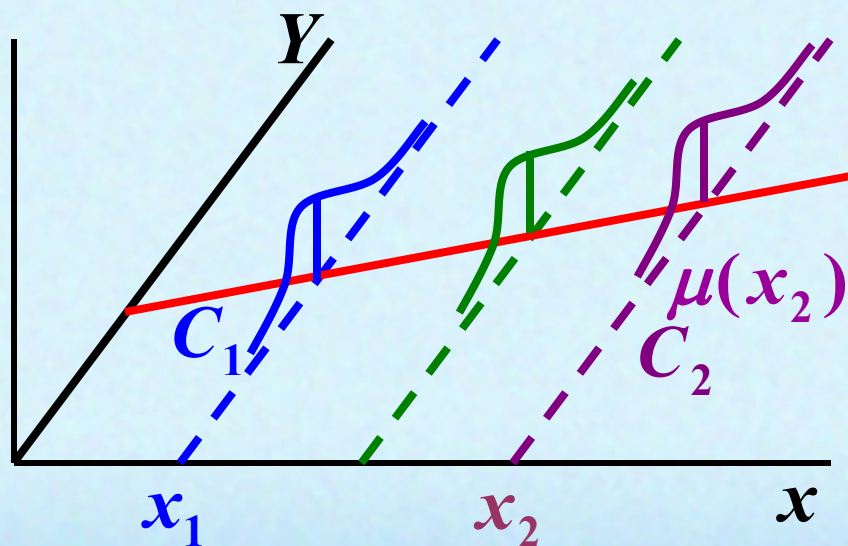
由于存在测量误差等原因,确定性关系在实际问题中往往通过相关关系表示出来;另一方面,当对事物内部规律了解得更加深刻时,相关关系也有可能转化为确定性关系.

回归分析—— 处理变量之间的相关关系的一种数学方法,它是最常用的数理统计方法.



二、一元线性回归的数学模型

设随机变量 Y (因变量) 与普通变量 x (自变量) 之间存在着相互关系. 由于 Y 是随机变量, 对于 x 的确定值, Y 有它的分布 (如图)



图中 C_1, C_2 分别是 x_1, x_2 处 Y 的概率密度曲线.

$F(y|x)$ 表示当 x 取确定的值 x 时, 所对应的 Y 的分布函数 $E(Y) = \mu(x)$.

因为对随机变量 η , 当 $c = E(\eta)$ 时, $E[(\eta - c)^2]$ 达到最小.

所以在一切 x 的函数中以回归函数 $\mu(x)$ 作为 Y 的近似, 均方误差 $E[(Y - \mu(x))^2]$ 为最小.

实际问题中的 $\mu(x)$ 一般未知.

问题的一般提法

对 x 的一组不完全相同的值 $x_1, x_2, \dots, x_n$, 设 $Y_1, Y_2, \dots, Y_n$ 分别是在 $x_1, x_2, \dots, x_n$ 处对 Y 的独立观察结果.

称 $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ 是一个样本.

对应的样本值记为 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

利用样本来估计 Y 关于 x 的回归函数 $\mu(x)$.
首先推测 $\mu(x)$ 的形式.

在一些问题中,可以由专业知识知道 $\mu(x)$ 的形式. 否则,可将每对观察值 (x_i, y_i) 在直角坐标系中描出它的相应的点, 这种图称为散点图.

求解步骤

1. 推测回归函数的形式

方法一 根据专业知识或者经验公式确定;

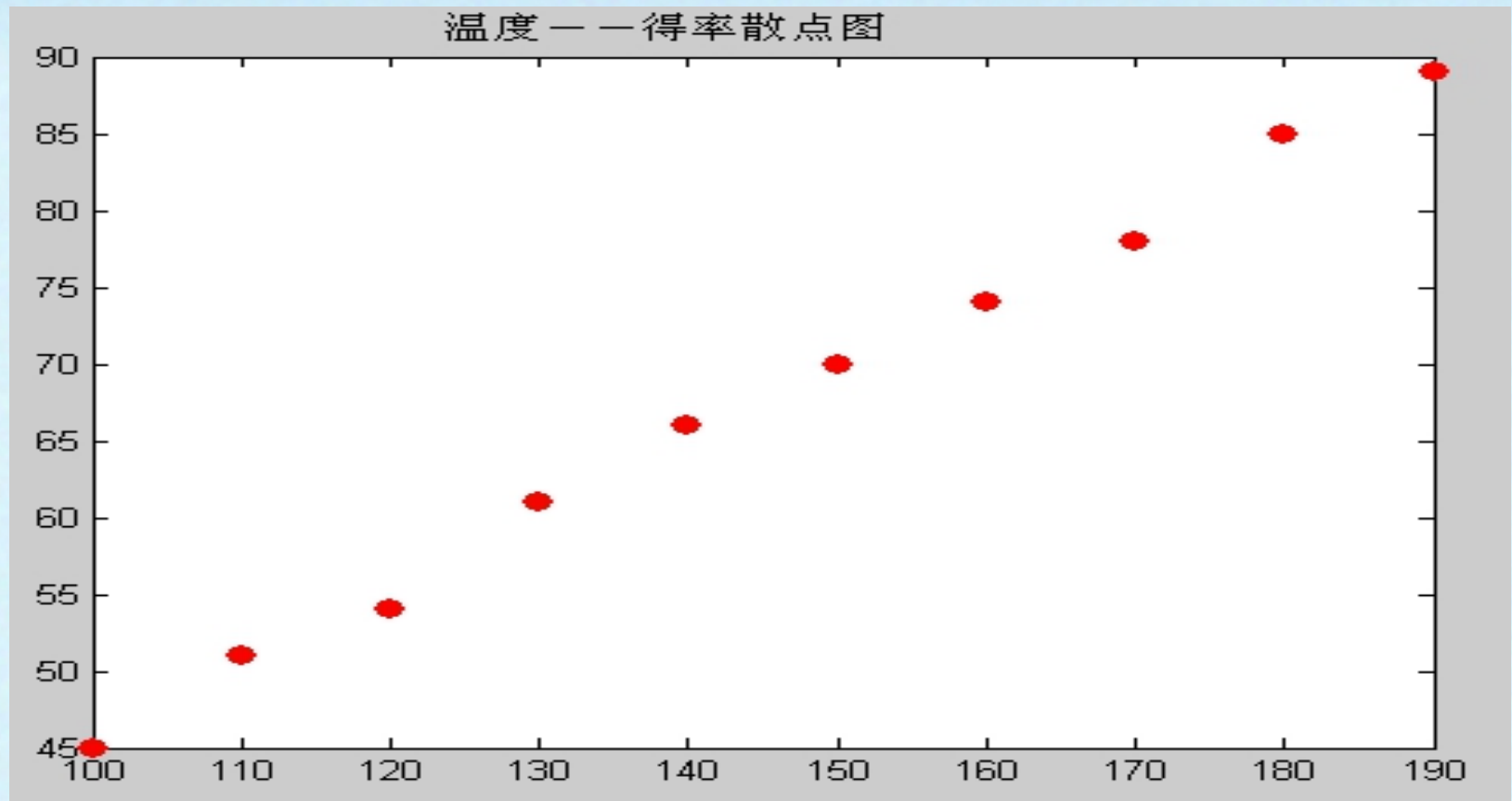
方法二 作散点图观察.

例1 为研究某一化学反应过程中, 温度 $x(^{\circ}C)$ 对产品得率 $Y(\%)$ 的影响, 测得数据如下.

温度 $x(^{\circ}C)$	100	110	120	130	140	150	160	170	180	190
得率 $Y(\%)$	45	51	54	61	66	70	74	78	85	89

这里自变量 x 是普通变量, Y 是随机变量.

画出散点图如下,



观察散点图， $\mu(x)$ 具有线性函数 $a + bx$ 的形式。

设 Y 关于 x 的回归函数为 $\mu(x)$.

利用样本来估计 $\mu(x)$ 的问题称为求 Y 关于 x 的回归问题 .

特别, 若 $\mu(x)$ 为线性函数: $\mu(x) = a + bx$,

此时估计 $\mu(x)$ 的问题称为求一元线性回归问题 .

2. 建立回归模型

$$\mu(x) = a + bx \quad \text{一元线性回归问题}$$

假设对于 x 的每一个值有 $Y \sim N(a + bx, \sigma^2)$,
 a, b, σ^2 都是不依赖于 x 的未知参数.

记 $\varepsilon = Y - (a + bx)$, 那么

$$Y = a + bx + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

a, b, σ^2 是不依赖于 x 的未知参数.

一元线性回归模型

x 的线性函数 随机误差

假设对于 x (在某个区间内)的每一个值

$$Y \sim N(a + bx, \sigma^2)$$

其中 a, b, σ^2 都是不依赖于 x 的未知参数. 记

$$\varepsilon = Y - (a + bx)$$

对 Y 作这样的正态假设, 相当于假设

$$Y = a + bx + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

其中未知参数 a, b 及 σ^2 都不依赖于 x . 上式称为一元线性回归模型, 其中 b 称为回归系数.

3. 未知参数 a, b 的估计

取 x 的 n 个不全相同的值 $x_1, x_2, \dots, x_n$ 做独立试验, 得到样本 $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$.

$Y_i = a + bx_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, 各 ε_i 相互独立.

于是 $Y_i \sim N(a + bx_i, \sigma^2)$, $i = 1, 2, \dots, n$. 由 $Y_1, Y_2, \dots, Y_n$ 的独立性, 知 $Y_1, Y_2, \dots, Y_n$ 的联合密度函数为

$$L = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2\sigma^2} (y_i - a - bx_i)^2 \right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - a - bx_i)^2\right].$$

用最大似然估计估计未知参数 a, b .

对于任意一组观察值 $y_1, y_2, \dots, y_n$,

L 就是样本的似然函数,

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - a - bx_i)^2\right] \text{要取最大值,}$$

只要上式右端方括弧中的平方和部分最小,

即只需函数

$$Q(a,b) = \sum_{i=1}^n (y_i - a - bx_i)^2 \text{取最小值.}$$

根据

$$\left. \begin{aligned} \frac{\partial Q}{\partial a} &= -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \\ \frac{\partial Q}{\partial b} &= -2 \sum_{i=1}^n (y_i - a - bx_i)x_i = 0 \end{aligned} \right\}$$

得方程组

$$\begin{cases} na + \left(\sum_{i=1}^n x_i\right)b = \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i\right)a + \left(\sum_{i=1}^n x_i^2\right)b = \sum_{i=1}^n x_i y_i \end{cases}$$

由于 x_i 不全相同, 方程组的系数行列式

$$\begin{vmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix} = n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 = n \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$$

$$\hat{b} = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{a} = \frac{1}{n} \sum_{i=1}^n y_i - \frac{\hat{b}}{n} \sum_{i=1}^n x_i = \bar{y} - \hat{b} \bar{x}$$

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

在得到 a, b 的估计 $\hat{a}, \hat{b}$ 后, 对于给定的 x ,

取 $\hat{a} + \hat{b}x$ 作为回归函数 $\mu(x) = a + bx$ 的估计, 即

$$\widehat{\mu(x)} = \hat{a} + \hat{b}x$$

称为 Y 关于 x 的经验回归函数, 记 $\hat{a} + \hat{b}x = \hat{y}$, 方程

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

称为 Y 关于 x 的经验回归方程, 简称**回归方程**, 其图形称为**回归直线**.

$$\hat{y} = \bar{y} + \hat{b}(x - \bar{x}),$$

对于样本值 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 回归直线通过散点图的几何中心 $(\bar{x}, \bar{y})$.

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2,$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2,$$

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

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

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \frac{1}{n} \sum_{i=1}^n y_i - \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \hat{b}.$$

例2 例1中的随机变量 Y 符合一元线性回归模型所述的条件,求 Y 关于 x 的线性回归方程.

	x	y	x^2	y^2	xy
	100	45	10000	2025	4500
	110	51	12100	2601	5610
	120	54	14400	2916	6480
	130	61	16900	3721	7930
	140	66	19600	4356	9240
	150	70	22500	4900	10500
	160	74	25600	5476	11840
	170	78	28900	6084	13260
	180	85	32400	7225	15300
	190	89	36100	7921	16910
Σ	1450	673	218500	47225	101570

$$S_{xx} = 218500 - \frac{1}{10} \times 1450^2 = 8250$$

$$S_{xy} = 101570 - \frac{1}{10} \times 1450 \times 673 = 3985$$

$$\hat{b} = S_{xy} / S_{xx} = 0.48303$$

$$\hat{a} = \frac{1}{10} \times 673 - \frac{1}{10} \times 1450 \times 0.48303 = -2.73935$$

回归直线方程 $\hat{y} = -2.73935 + 0.48303x$

或 $\hat{y} = 67.3 + 0.48303(x - 145)$

当 x (温度)增加一单位时, y (得率)平均增加0.48303个单位。

例1的回归实验

例1回归实验 - Microsoft Excel

文件 开始 插入 页面布局 公式 数据 审阅 视图 Acrobat

获取外部数据 全部刷新 排序 筛选 分列 删除 重复项 分级显示

数据 数据工具 数据分析

数据 I17

	A	B
1	100	45
2	110	51
3	120	54
4	130	61
5	140	66
6	150	70
7	160	74
8	170	78
9	180	85
10	190	89
11		
12		

分析工具 (A)

- 指数平滑
- F-检验 双样本方差
- 傅利叶分析
- 直方图
- 移动平均
- 随机数发生器
- 排位与百分位排位
- 回归
- 抽样
- t-检验：平均值的成对二样本分析

确定 取消 帮助 (H)

Sheet1 Sheet2 Sheet3

就绪 100%

回归



输入

Y 值输入区域 (Y):

\$B\$1:\$B\$10



X 值输入区域 (X):

\$A\$1:\$A\$10



☐ 标志 (L)

☐ 常数为零 (Z)

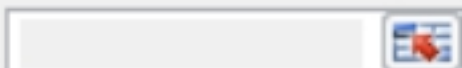
☒ 置信度 (F)

95

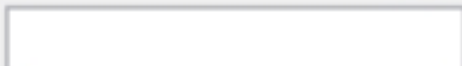
%

输出选项

☐ 输出区域 (O):



☒ 新工作表组 (P):



☐ 新工作簿 (W)

残差

☐ 残差 (R)

☒ 残差图 (Q)

☐ 标准残差 (T)

☒ 线性拟合图 (I)

正态分布

☐ 正态概率图 (N)

确定

取消

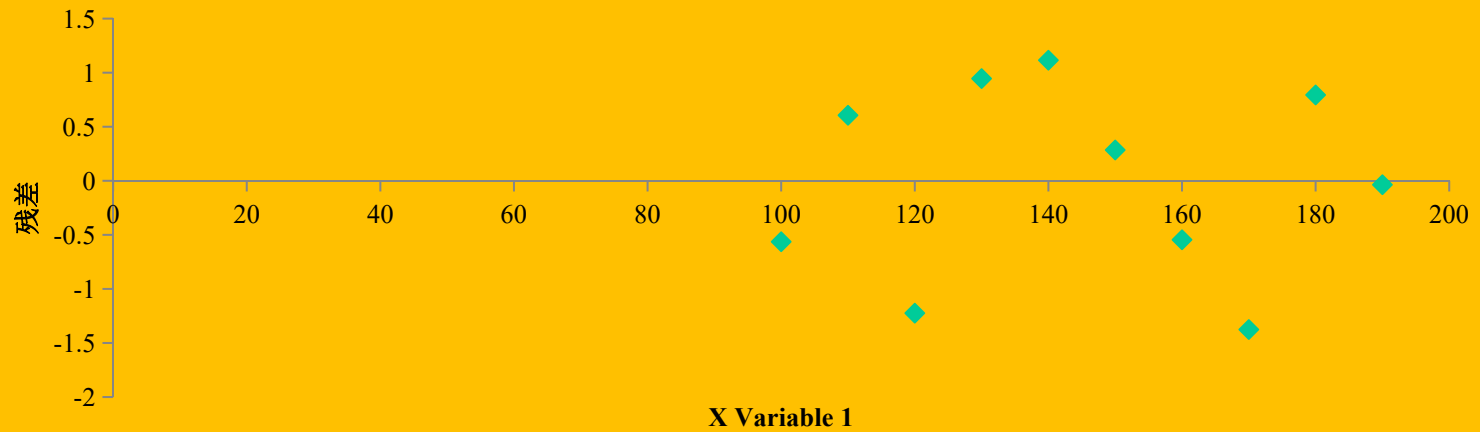
帮助 (H)

SUMMARY OUTPUT						
回归统计						
Multiple	0.998129					
R Square	0.996261					
Adjusted	0.995794					
标准误差	0.950279					
观测值	10					
方差分析						
	df	SS	MS	F	F_p	
回归分析	1	1924.876	1924.876	2131.574	5.35253E-11	
残差	8	7.224242	0.90303			
总计	9	1932.1				
Coefficients		标准误差	t Stat	P-value	下限 95.0%	上限 95.0%
Intercept	-2.73939	1.5465	-1.77135	0.11445	-6.305629204	0.826841
X Variable	0.48303	0.010462	46.16897	5.35E-11	0.458904362	0.507156

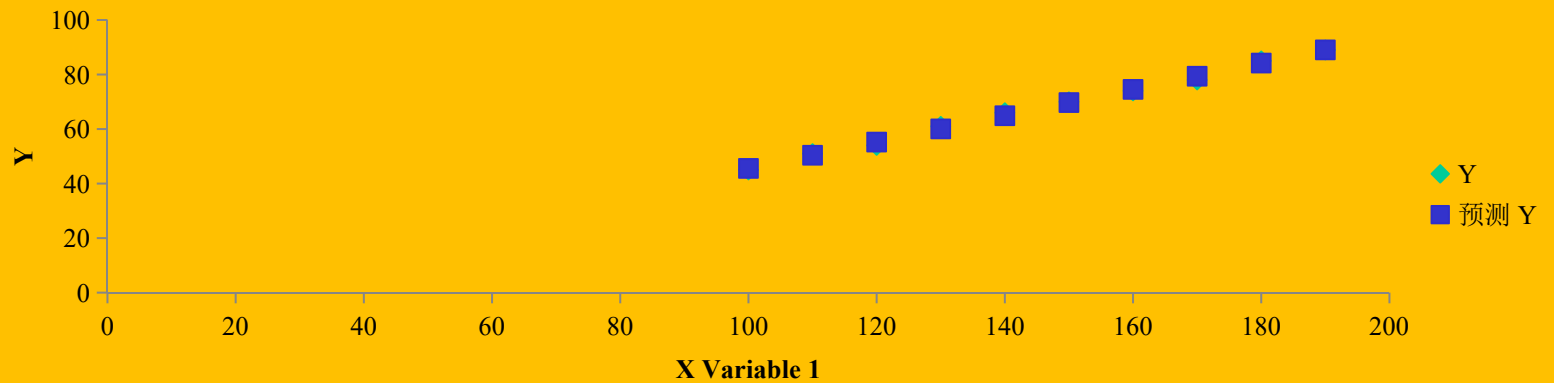
$$\hat{y} = -2.73935 + 0.48303x$$

9.3 一元线性回归

X Variable 1 Residual Plot



X Variable 1 Line Fit Plot



四、小结与思考

1. 回归分析的任务

研究变量之间的相关关系

2. 一元线性回归的步骤

- (1) 推测回归函数;
- (2) 建立回归模型;
- (3) 估计未知参数;
- (4) 通过实验进行参数求解和假设检验;

3. 思考:

- (1) 参数的估计及估计的效果?
- (2) 模型的应用 (预测)

谢 谢！