第九章　回归分析

§9.1 相关关系与回归分析

§9.2　一元回归分析

§9.1 回归分析与回归方程

- 相关关系与回归分析
- 回归方程

一. 相关关系与回归分析

在现实世界中存在大量的变量, 它们有相互依存、相互制约的关系，一般分为两类：**确定性关系与非确定性关系.**

例1、正方形的面积S与边长X之间的关系

例2、人的体重W与身高H之间的关系

例3、农作物产量Y与降雨量X_1，氮、磷、钾的施肥量X_2、X_3、X_4之间的关系

思考：这三个例子各有什么特点?

例1 是确定关系

例2 和例3 是相关关系（不确定关系）

(例3中X_2、X_3、X_4可精确度量，而X_1不能)

回归分析

相关关系的基本特征是一个变量不能依据其他有关变量的数值精确地求出其数值。

我们可以根据大量统计数据，找出变量之间在数量变化方面的规律，这种统计规律称**回归关系**。

表示这种规律的数学公式称为**回归方程**。

有关回归关系的计算方法和理论称为**回归分析**。

TIPS

“回归”一词的由来

现实中常需要研究随机变量间的关系.

问题 如何描述各变量间的关系?

将作为考察目标的变量称为**因变量**(记为Y), 而将影响它的各个变量称为**自变量**或**可控变量**, 记为$(X_1, X_2, \cdots X_k)$

由于统计相关的随机性，考察变量的关系是：当自变量取某个确定值时，因变量所有可能出现的对应值的平均值。

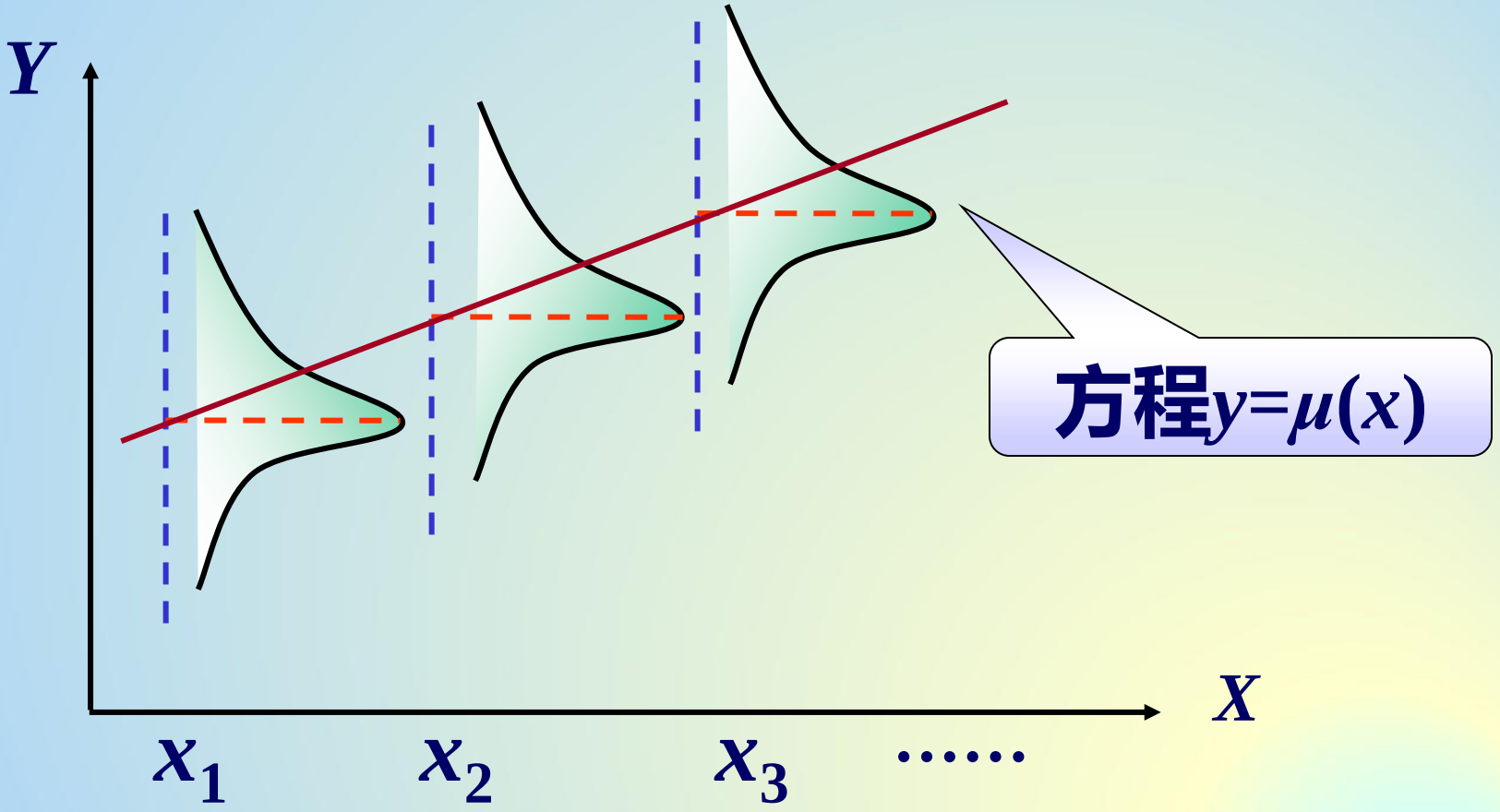
特别对因变量Y与单个自变量X，在“$X=x$”时，Y的条件数学期望为

$$\mu(x)=E(Y|X=x)=\int_{-\infty}^{+\infty}yf_{Y|X}(y|x)dy$$

1) $\mu(x)$可理解为在“$X=x$”的条件下，随机变量Y取值最集中的点；

2）方程$y=\mu(x)$描述了Y与X间非确定性的相关关系.

对于X的不同取值x_1，x_2，…，x_n



定义9.1.1 称

$$\mu(x_1, x_2, \cdots, x_k) = E(Y \mid X_1 = x_1, X_2 = x_2, \cdots X_k = x_k)$$

为Y关于X_1，X_2，$\cdots$ X_k的**回归函数，**

方程 $y = \mu(x_1, x_2, \cdots, x_k)$

称为Y对 X_1，X_2，$\cdots$ X_k的**回归方程.**

注 回归函数是确定性的函数.

回归分析即以回归函数为基础处理相关关系的一种方法.

3.回归模型的引进

若Y关于X_1，X_2，$\cdots$ X_k的回归方程为

$$y=\mu(x_1,x_2,\cdots,x_k)$$

由变量间不存在确定函数关系，引入随机误差

得数学模型：

$$Y=\mu(x_1,x_2,\cdots,x_k)+\varepsilon$$

有 $\varepsilon=Y-\mu(x_1,x_2,\cdots,x_k)$

ε可视为随机误差，通常要求：

其它未知的、未考虑的因素以及随机因素的影响所产生.

1) $E(\varepsilon)=0$;

2) $D(\varepsilon)=E(\varepsilon^2)=\sigma^2$ 尽可能小.

注意到 $\sigma^2 = E[Y-\mu(x_1,x_2,\cdots,x_n)]^2$

σ^2 是用回归函数近似因变量Y产生的均方误差.

建立模型涉及三个问题：

1) 确定对因变量Y影响显著的自变量；

2) 确定回归函数$\mu(x)$的类型；

3) 对参数进行估计.

} 本章内容

二. 回归函数类型的确定

实际问题中，回归函数形式通常未知.

回归分析的**基本思想**：

根据自变量X_1，X_2，$\cdots$ X_k与因变量Y的观察值去**估计**回归函数.

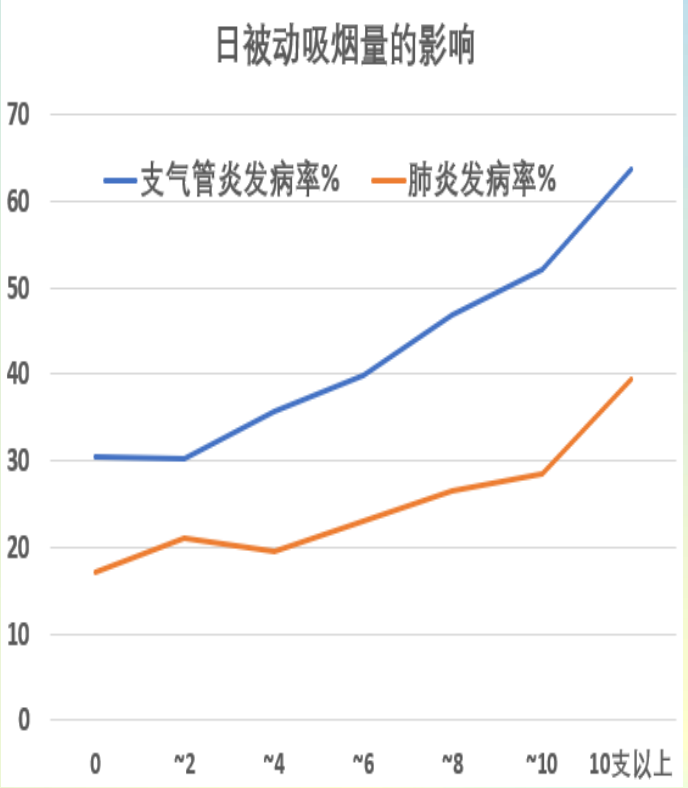
本节仅讨论最简单的情形：可控变量关于单个因变量的回归函数 $\mu(x)=E(Y\mid X=x)$.

为估计回归函数，可依据问题的背景，确定或**假定回归函数的形式.**

常通过分析数据散点图（或连线图）获得对变量间相关关系的初步认识.

例9.1.1 日被动吸烟量对3~6岁儿童呼吸系统疾病发病率的影响

日被动吸烟量(克/日)	调查人数	支气管炎发病人数	支气管炎发病率%	肺炎发病人数	肺炎发病率%
0	239	73	30.5	41	17.2
~2	142	43	30.3	30	21.1
~4	143	51	35.7	28	19.6
~6	148	59	39.9	34	23
~8	49	23	46.9	13	26.5
~10	67	35	52.2	19	28.4
10支以上	33	21	63.6	13	39.4
合计	821	305	37.1	178	21.7



例9.1.2 施肥效果分析

某地区作物生长所需的营养素主要是氮(N)、钾(K)、磷(P).某作物研究所在某地区对土豆做了一定数量的实验,实验数据如下列表所示,其中ha表示公顷,试分析施肥量与土豆产量之间关系.

回归分析

N

施肥量 (kg/ha)	产量 (t/ha)
0	15.18
34	21.36
67	25.72
101	32.29
135	34.03
202	39.45
259	43.15
336	43.46
404	40.83
471	30.75

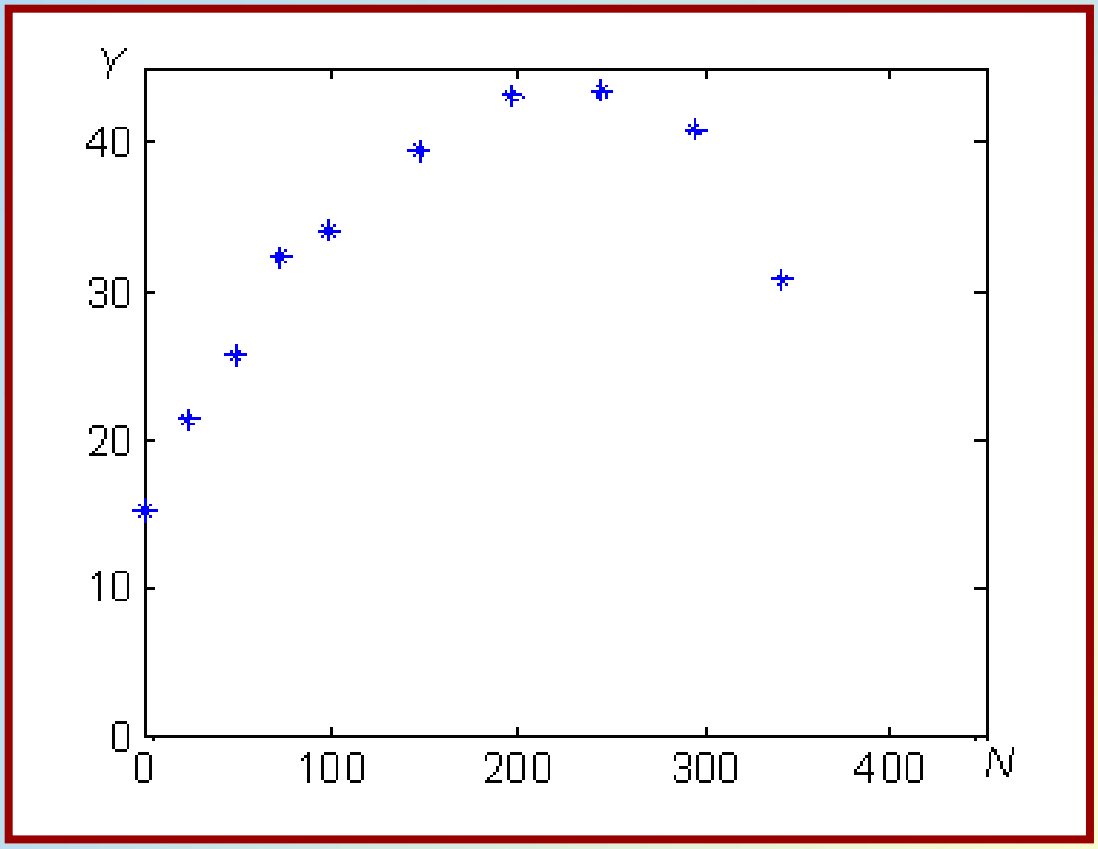
P

施肥量 (kg/ha)	产量 (t/ha)
0	34.46
24	32.47
49	36.06
73	37.96
98	41.04
147	40.09
196	41.26
245	42.17
294	40.36
342	42.73

K

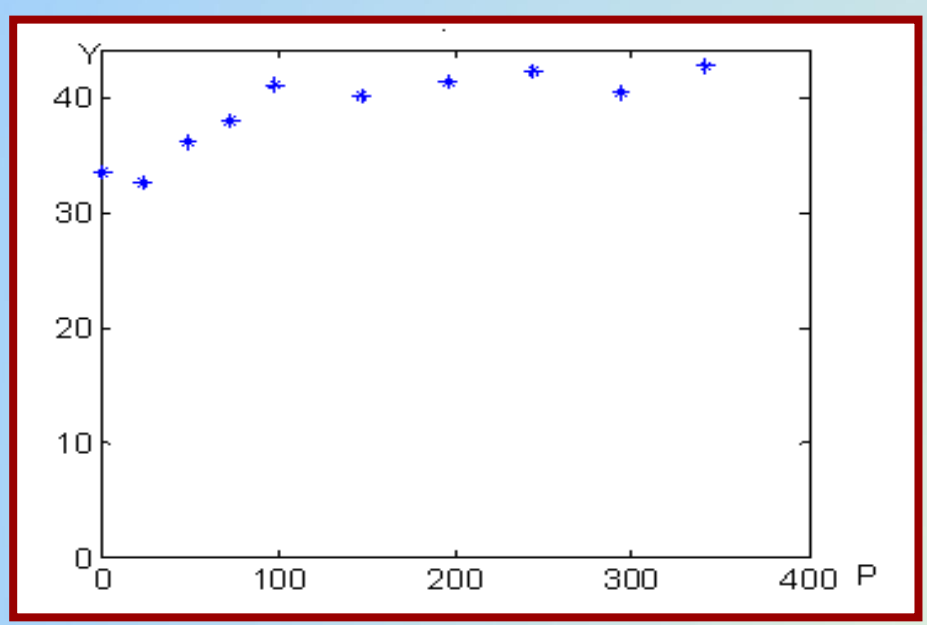
施肥量 (kg/ha)	产量 (t/ha)
0	18.98
47	27.35
93	34.86
140	39.92
186	38.44
279	37.73
372	38.43
465	43.87
558	42.77
651	46.22

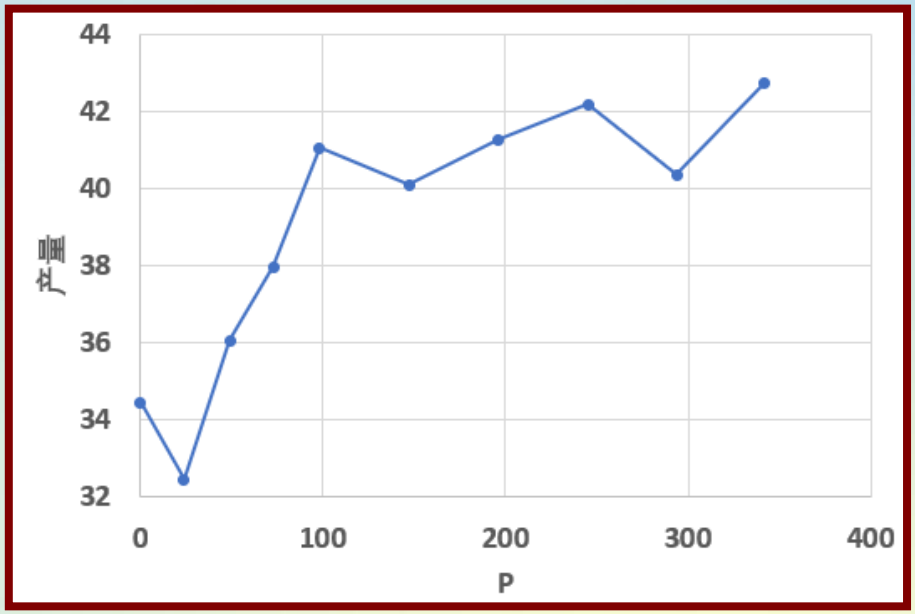
土豆产量—氮肥量数据散布图



$$y = \hat{\mu}(x) = b_0 + b_1 x + b_2 x^2$$

土豆产量—磷肥量数据散布图





可选 $y = \hat{\mu}(x) = \dfrac{1}{a + be^{-x}}, x \geq 0$

思考 是否能由数据散布图完全确定回归函数?

结论 仅是初步感性的认识，还需进行检验.