



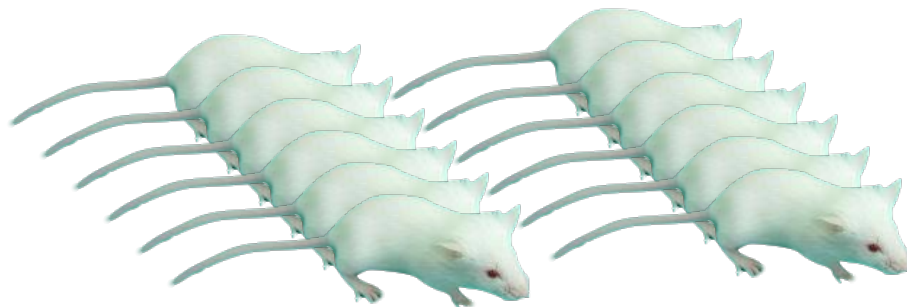
糖尿病患者



某人群



饲料



12只大白鼠



婚姻状况



生育数量

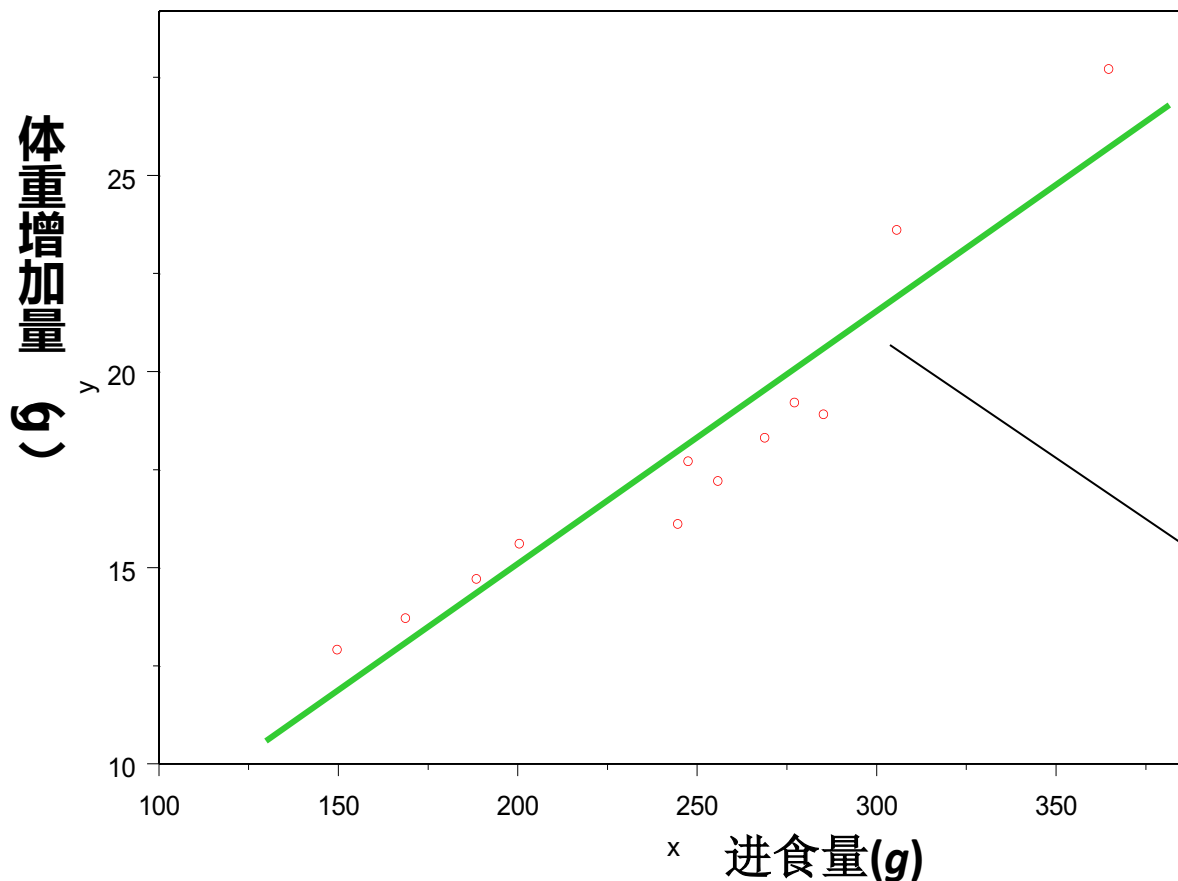
案例：

用某饲料喂养**12**只大白鼠，得出大鼠进食量与体重增加量如表**1**。试分析大鼠的进食量和体重增加量之间的回归关系。

表1 12只大白鼠进食量 (g) 与体重增加量 (g)

序号(1)	进食量 X (2)	体重增加值 Y (3)
1	305.7	23.6
2	188.6	14.7
3	277.2	19.2
4	364.8	27.7
5	285.3	18.9
6	244.7	16.1
7	255.7	17.2
8	149.8	12.9
9	268.9	18.3
10	247.6	17.7
11	168.8	13.7
12	200.6	15.6
合计	2957.9 ΣX	215.6 ΣY

案例中：



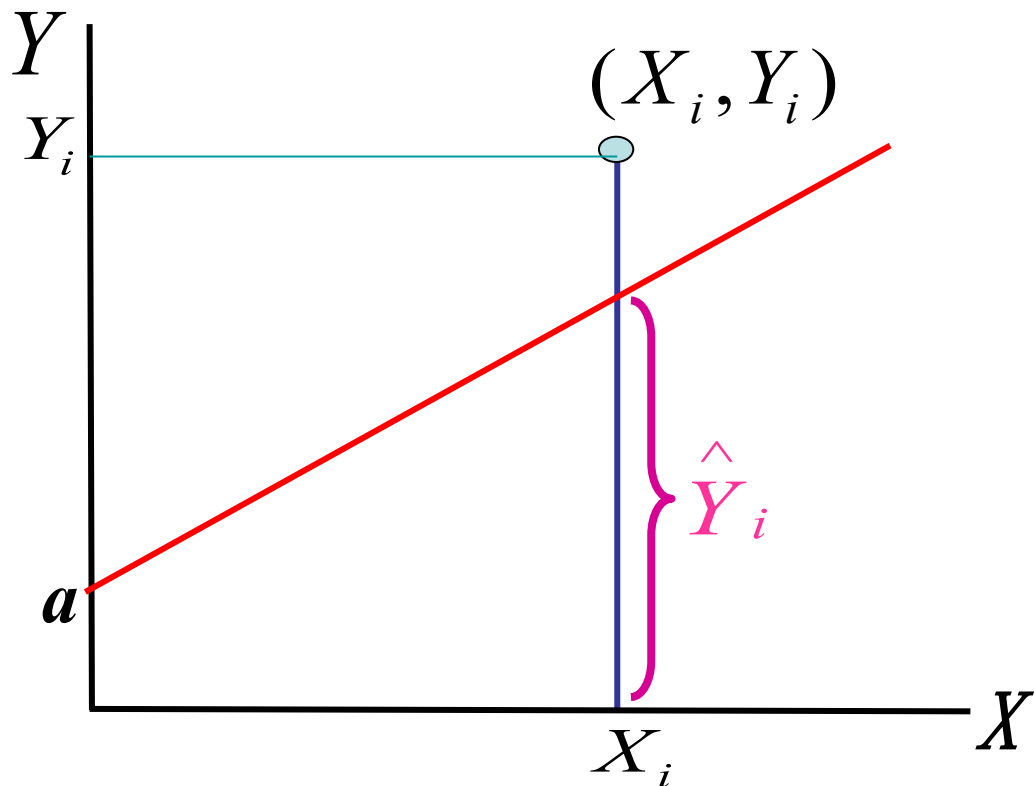
直线回归 /
简单回归
**linear regression /
simple regression**

直线回归方程
**linear regression
equation**

12只大白鼠的进食量(g)与体重增加量(g)散点图

一元线性回归方程： $Y = \alpha + \beta X + \varepsilon$

模型系数 残差



一般表达式：

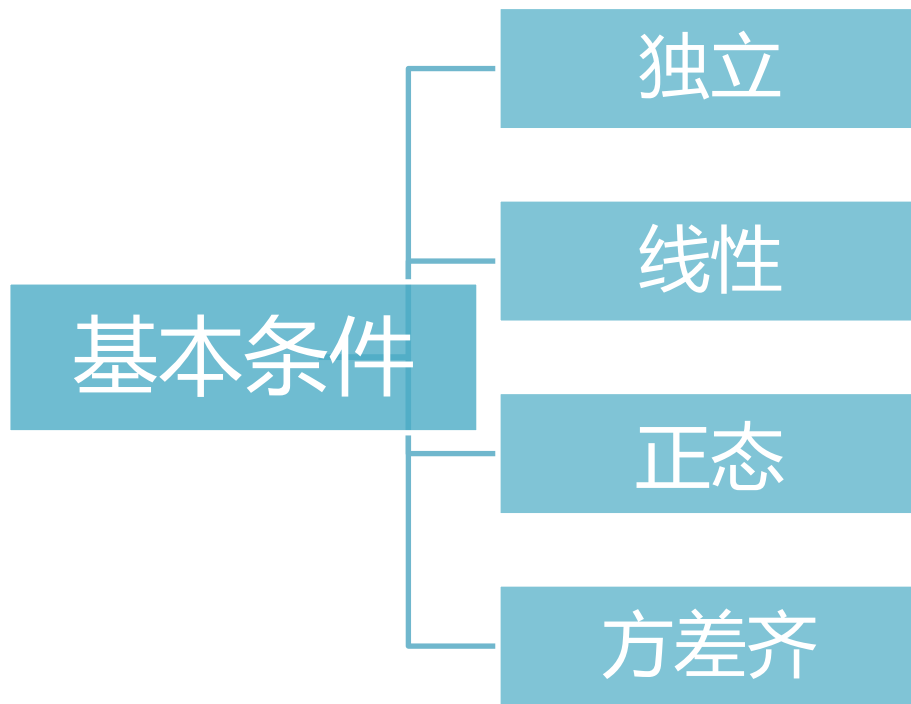
$$\hat{Y} = a + bX$$

应变量 Y 的估计值
估计值
自变量

a、b是 α 、 β 的估计值。

a：截距(intercept), 直线与Y轴交点的纵坐标。

b：斜率(slope), 样本回归系数(regression coefficient)
意义： X 每改变一个单位， Y 平均改变 b 个单位。



了解：均值、标准差、最大值、最小值、正态分布情况
观察：质量、缺少值、异常值

最小二乘法(least square method)

使因变量的观察值(Y)与估计值(\hat{Y})之间的离差平方和达到最小来求得 a 和 b 的方法。即：

$$Q(a, b) = \sum_{i=1}^n (Y_i - \hat{Y})^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2 = \sum_{i=1}^n e_i^2 \rightarrow \min$$

该方程组有唯一的一组解，即：

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{l_{XY}}{l_{XX}}$$

$$a = \bar{Y} - b\bar{X}$$

$$\sum (Y_i - \hat{Y}_i) = 0$$

过点 (\bar{X}, \bar{Y})

表1 12只大白鼠的进食量(g)与体重增加量(g)

序号(1)	进食量X (2)	体重增加值Y (3)	X^2 (4)	Y^2 (5)	XY (6)
1	305.7	23.6	93452.49	556.96	7214.52
2	188.6	14.7	35569.96	216.09	2772.42
3	277.2	19.2	76839.84	368.64	5322.24
4	364.8	27.7	133079.04	767.29	10104.96
5	285.3	18.9	81396.09	357.21	5392.17
6	244.7	16.1	59878.09	259.21	3939.67
7			65484.81	295.84	4401.48
8			22440.04	166.41	1932.42
9	268.9	18.3	72307.21	334.89	4920.87
10	247.6	17.7	61305.76	313.29	4382.52
11	168.8	13.7	28493.44	187.69	2312.56
12	200.6	15.6	40240.36	243.36	3129.36
合计	2957.9	215.6	770487.13	4066.9	55825.2
	ΣX	ΣY	ΣX^2	ΣY^2	ΣXY

$$l_{XX} = \sum X^2 - \frac{(\sum X)^2}{n} = 770487.1 - \frac{2957.9^2}{12} = 41398.4$$

$$l_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 4066.9 - \frac{215.6^2}{12} = 193.3$$

$$l_{XY} = \sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$
$$= 55825.2 - \frac{2957.9 \times 215.6}{12} = 2681.6$$

从而，回归系数 $b = \frac{l_{XY}}{l_{XX}} = \frac{2681.6}{41389.4} = 0.0648$

$$\text{截距 } a = \bar{Y} - b\bar{X} = 17.97 - 0.0648 \times 246.49 = 2.00$$

例1资料的回归方程： $\hat{Y} = 2.00 + 0.0648X$

X 和 Y 是否存在线性回归关系

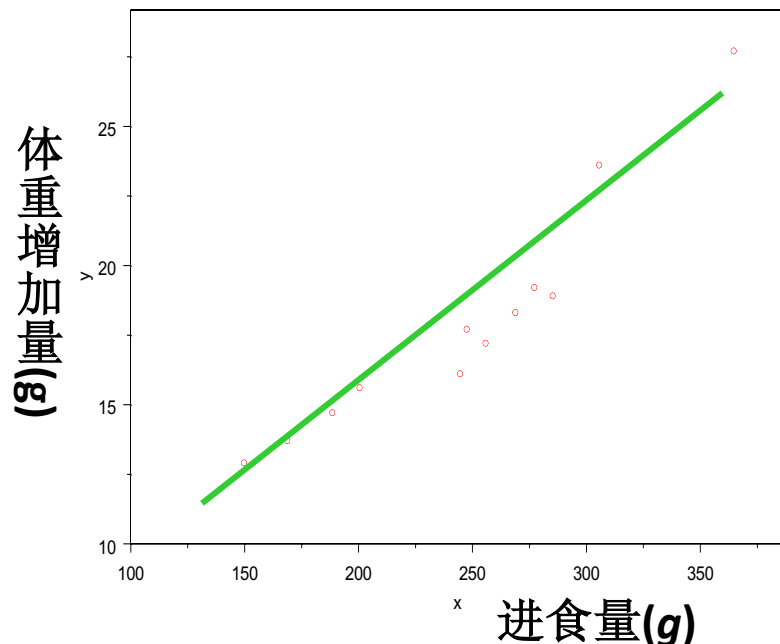


图1 12只大白鼠的进食量(g)与体重增加量(g)散点图

$$\hat{Y} = a + bX$$

对 β 是否为0进行假设检验

样本回归系数

不存在线性关系 $\rightarrow \beta=0$

$b \neq 0$ 的原因:

- 存在回归关系，总体回归系数 $\beta \neq 0$
- 由于抽样误差引起，总体回归系数 $\beta=0$

回归系数的t 检验

解：1. 提出检验假设，确定显著性水平

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0 \quad \alpha = 0.05$$

$$\sqrt{F} = \sqrt{88.6} = 9.41 \approx t$$

2. 计算统计量

$$t = \frac{b - 0}{S_b} = \frac{0.0648}{0.00688} = 9.42, \quad \nu = 10$$

3. 确定P值，下结论 查t界值表， $P < 0.001$ ，

按 $\alpha = 0.05$ 的检验水准拒绝 H_0 ，接受 H_1 ，故可认为

体重的增加量与进食量之间有直线关系。

例题

样本回归系数 $b=0.0648$ ，估计总体回归系数 b 的95%可信区间。

解：

$$S_b=0.00688, \quad df=12-2=10$$

查 t 界值表，得 $t_{0.05/2,10}=2.228$ ，故 β 的95%可信区间是

$$(0.0648-2.228 \times 0.00688, \quad 0.0648+2.228 \times 0.00688)$$

$$=(0.0495, \quad 0.0801)$$