第二节　回归分析

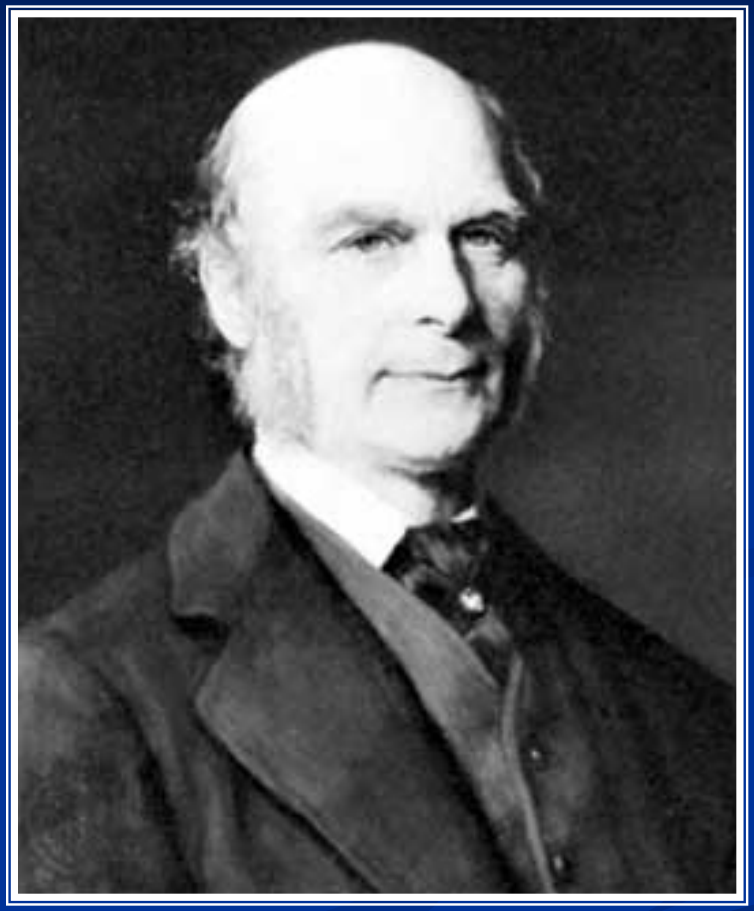
相关系数只能说明现象间相关关系的方向和程度，关系密切与否，但不能说明一个现象发生一定量的变化，另一个现象一般也会发生多大的变化。

回归分析是研究变量与变量之间相关关系的一种统计推断方法。它是在试验观测数据的基础上，寻找被随机性掩盖了的变量之间的相互依存的关系，以一种确定的函数关系去近似替代比较复杂的相关关系。

弗兰西斯·高尔顿于1822年生于英格兰，与达尔文是表兄弟关系，他从小智力超常、聪颖过人，被誉为神童，是著名的优生学家、心理学家，差异心理学之父，也是心理测量学上生理计量法的创始人，享年89岁。

高尔顿一生在统计学方面贡献很多，首次引入了“Regression回归”一词，第一次使用了相关系数 (correlation coefficient) 的概念，并采用字母“ *r* ”来表示。

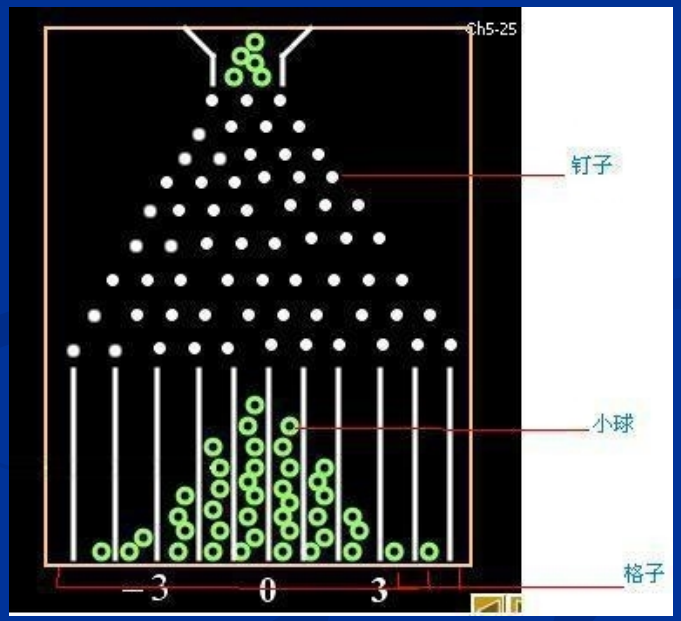
高尔顿设计的用来研究随机现象的高尔顿钉板模型，更是被广泛用来描述正态分布的经典例子。



高尔顿（ Francis Galton）
1822－1911

高尔顿(Galton)钉板试验

试验模型如下所示：



自上端放入一小球，任其自由下落，在下落过程中当小球碰到钉子时，从左边落下与从右边落下的机会相等. 碰到下一排钉子时又是如此, 最后落入底板中的某一格子. 因此，任意放入一球，则此球落入哪一个格子，预先难以确定.但是如果放入大量小球，则其最后所呈现的曲线 几乎总是一样的.

,

一、回归分析的涵义

1、高尔顿试验

“Regression(回归)”一词是由英国著名人类学家、气象学家和统计学家高尔顿于1885年在其《身高遗传中的平庸回归》一文中首次引入的，他在研究身高与遗传之间的联系时，观察了1078对夫妇的二人的平均身高 X 以及其一个成年后代的身高 Y，从中发现在直角坐标系下，二者之间的关系近乎是一条直线，并且得到如下数学关系：

$$\hat{Y} = 33.73 + 0.516\,X$$

高尔顿的结论是：

结果解释

父辈平均身高每增加或减少一个单位，其成年后代的身高平均增加或减少0.516 个单位。

从人类遗传上来说，父母个子高这一基因会遗传给他们的后代，导致产生高个子的下一代，但子代的身高并不会象其父辈，出现越来越高的现象，而是趋向于比他们父辈身高更加平均的水平。
高尔顿将人类这种遗传现象称为“ 回归 ”。人类也正是由于这种回归，才能生生不息的繁衍下去。

2、何谓回归分析

回归分析-----是对具有相关关系(显著相关以上相关)的两个或两个以上的变量之间所具有的变化规律进行拟合，确立一个相应的数学表达式(经验公式)，通过一个或多个变量的变化去解释另一变量变化的方法，以便从定量的角度由已知量推测未知量，为估算预测或控制提供重要依据。

简单的说，回归分析就是一种处理具有相关关系的变量与变量之间关系的数学方法与工具。

3、回归分析的内容和步骤

回归分析需要研究和解决的问题主要有以下几方面：

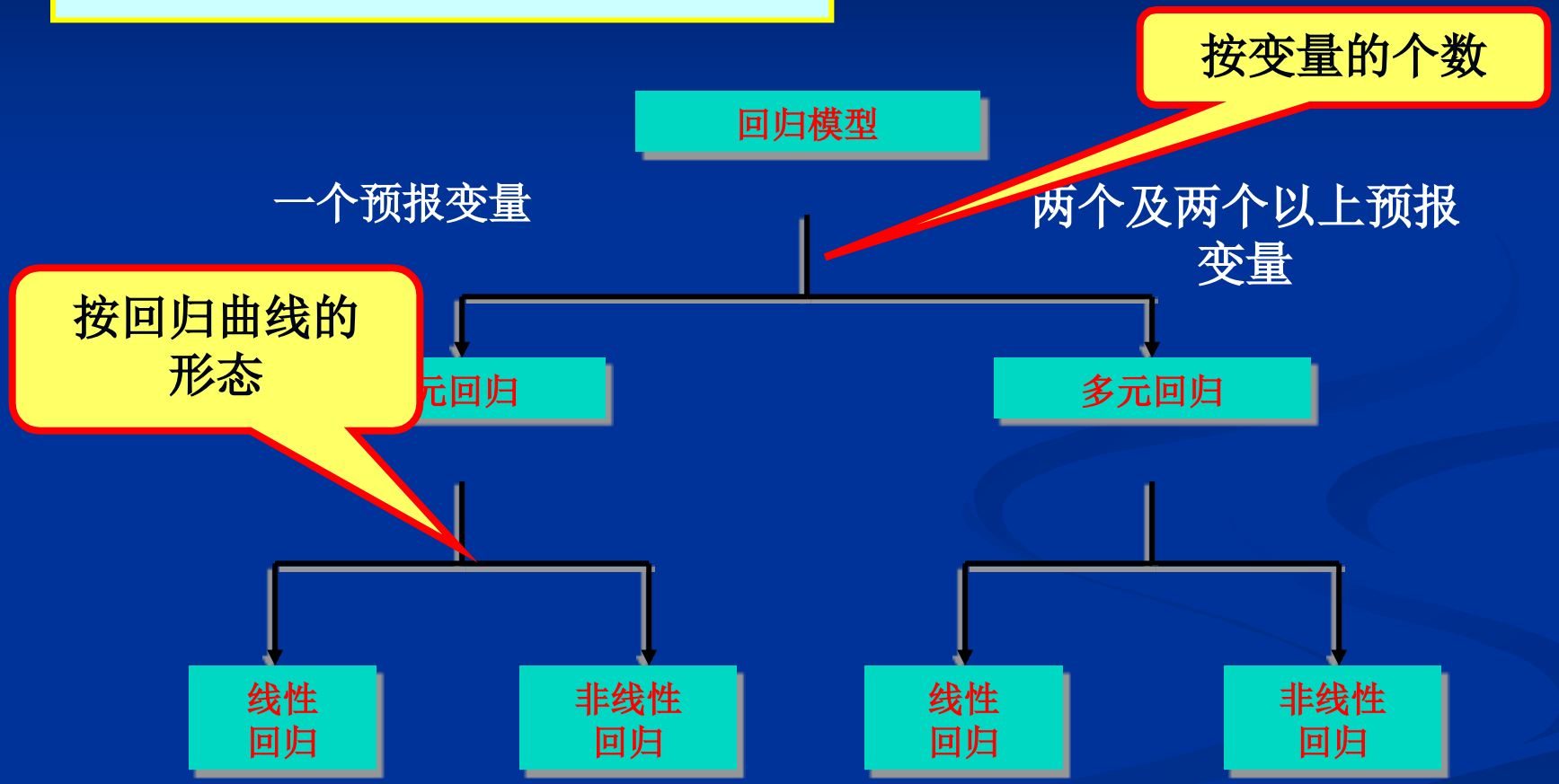
(1)根据理论和对实际问题的分析判断, 区分自变量(即解释变量或预报变量)和因变量(即被解释变量或响应变量)。

(2)从一组试验数据出发，判断二者之间是否存在相关关系, 如果存在的话，设法找出其合适的数学表达式(即回归模型)用来描述变量之间的内在联系。

(3)对建立的回归模型可信程度进行统计检验和推断, 并从影响因变量的诸多自变量中找出影响显著或不显著的变量。

(4)依据回归模型, 通过自变量的取值来预测或控制因变量的取值, 并给出这种预测或控制的精确程度。

4、回归模型的分类


按变量的个数
回归模型
一个预报变量
两个及两个以上预报变量
按回归曲线的形态
元回归
多元回归
线性回归
非线性回归
线性回归
非线性回归

5、回归数学模型

(1) 一般回归模型

$$\begin{cases} Y = f(X_1, X_2, \mathrm{L}\ , X_k) + \varepsilon \\ E(\varepsilon) = 0,\ D(X_i) = \sigma^2 \end{cases}, \text{ 其中: } R.V.X_i,\ i = 1, 2, \mathrm{L}\ , k$$

式中：① ε 是不可观察的随机误差, 其分布是与控制变量 $X_1, X_2, \ldots, X_k$ 无关的随机变量。

② $f\ (X_1, X_2, \ldots, X_k)$ 称为回归函数,它是一个未知的多元函数。

③ X_i 称为控制变量(解释变量、预报变量), Y 称为响应变量。

④ $E(Y) = f(X_1, X_2, \mathrm{L}\ , X_k) + E(\varepsilon) = f(X_1, X_2, \mathrm{L}\ , X_k)$

(2) 线性回归模型

如果响应变量Y和控制变量$X_1, X_2, \ldots, X_k$呈现线性相关关系的情形，即

$$f(X_1, X_2, \mathrm{L}, X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathrm{L} + \beta_n X_n$$

则称回归模型

$$\begin{cases} Y = \beta_0 + \beta_1 X_1 + \mathrm{L} + \beta_n X_n + \varepsilon \\ E(Y) = \beta_0 + \beta_1 X_1 + \mathrm{L} + \beta_n X_n \\ E(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2 < +\infty \text{(未知)} \end{cases}$$

为n元线性回归模型。

注：当线性回归模型只有一个控制变量时，称为一元线性回归模型，有多个控制变量时称为多元线性回归模型。

6、回归数学模型

	相关分析	回归分析
区别	变量之间地位对等, 无主从之分。	有因变量(处在被解释的地位)和自变量(控制变量, 用于预测因变量的变化)之分。
	涉及到的变量都是随机变量。	因变量是随机变量，自变量可以是随机变量, 也可以是非随机的确定变量。
	主要描述变量之间相关关系的密切程度与方向。	不仅可以揭示自变量对因变量的影响大小, 还可利用回归模型进行预测和控制。
联系	回归分析和相关分析的理论和方法具有一致性，相关系数和回归系数的方向一致, 可以互相推算。	
	回归分析和相关分析是互为补充、密切联系的两个不同概念，回归分析是建立在相关分析基础上的。相关系数的大小决定是否需要进行回归分析。同时相关系数还是检验回归模型效果好坏的标准。	
	无相关就无回归, 相关程度越高，回归效果越好, 所得到的统计推断、预测, 控制等精确度就越高。	
注	回归分析并不是建立变量之间的一个必然因果关系的过程，建立的回归方程只表明：变量是如何或者是以怎样的程度彼此相互联系在一起的。	

二、一元线性回归分析

1、一元线性回归数学模型

$$\begin{cases} Y = a + bX + \varepsilon \\ E(Y) = a + bX \\ E(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2 \end{cases}$$

ε 是是随机因素，是不可观察的随机变量，是许多不可控制或不了解的随机因素的总和，且满足

$$E(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2$$

其中：

X 是可控(或可观察)的非随机变量，常称为自变量，或预报变量。

a, b, σ^2都是未知参数，且都不依赖于X。

任务：估计线性回归方程中的未知参数

因为具有显著相关关系 y 不仅受 x 影响，还受其它因素影响，因此，x、y 形成的点不一定全在直线上，而是分在直线上下波动，呈现线性相关的趋势，所以需要在这些分散的相关点之间配合一条最合适的直线，用来模拟两变量之间具体的变动关系

2、未知参数的估计

建立线性回归方程的关键是估计未知参数, 具体步骤如下：

(1) 采集样本数据 ($x_1, x_2, \cdots, x_n$)

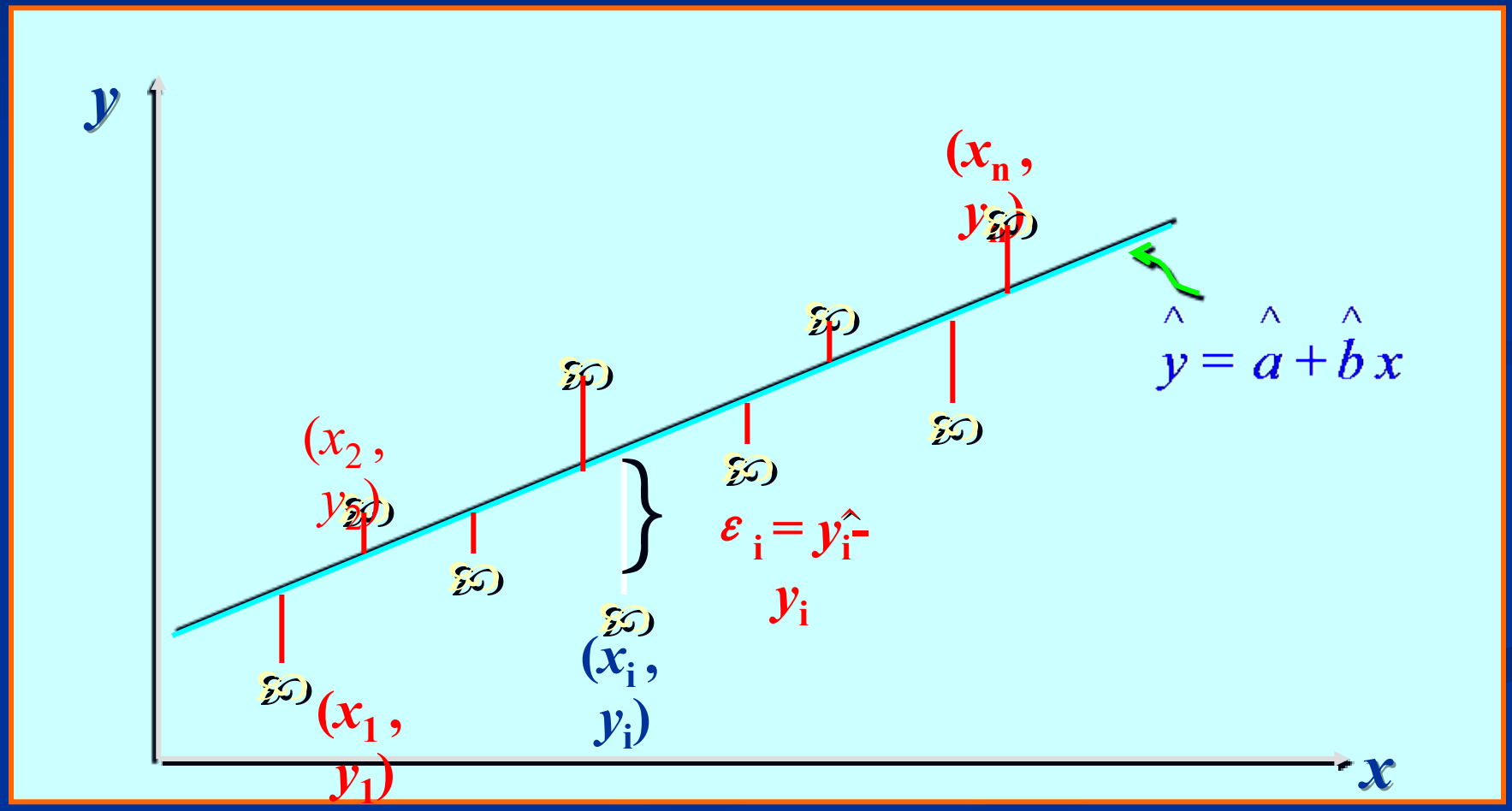
通常借助于 n 次独立试验或观察, 获得试验数据, 并利用其相应的观测值去估计

$$(x_1, y_1), (x_2, y_2), \mathrm{L}, (x_n, y_n)$$

$$\begin{cases} y_i = a + bx_i + \varepsilon_i \\ E(y_i) = a + bx_i \\ E(\varepsilon_i) = 0,\ D(\varepsilon_i) = \sigma^2 \end{cases}$$

, 其中: ε_i 是相互独立, $i = 1, 2, \mathrm{L}, n$

(2) 几何图示



(3) 借助于最小二乘原理进行统计估计

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, 即

$$Q(\hat{a},\hat{b})=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2=\sum_{i=1}^{n}(y_i-\hat{y})^2=\sum_{i=1}^{n}\varepsilon_i^2=Min$$

② 利用多元函数极值优化

$$\begin{cases}\dfrac{\partial Q(a,b)}{\partial a}=0\\[2ex]\dfrac{\partial Q(a,b)}{\partial b}=0\end{cases}$$

正规方程组

$$\begin{cases}n\,a+n\,\bar{x}\,b=n\,\bar{y}\\[2ex]n\,a\,\bar{x}+b\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i\end{cases}$$

其中: $\bar{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i,\ \bar{y}=\dfrac{1}{n}\sum_{i=1}^{n}y_i$

③ 根据最小二乘原理, 解正规方程组, 可求得估计值：

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

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

回归系数

由此可得二个变量的一元线性回归方程：

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

回归系数意义：当自变量增加或减少一个单位, 因变量将增加或减少 $\hat{b}$ 个单位。

用最小二乘法拟合出的这个线性方程(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都小。

引入记号：

$$L_{xx}=\sum_{i=1}^{n}x_i^2-n(\overline{x})^2,\quad L_{yy}=\sum_{i=1}^{n}y_i^2-n(\overline{y})^2,\quad L_{xy}=\sum_{i=1}^{n}x_i\,y_i-n\,\overline{x}\,\overline{y}$$

则回归系数可简化为：

$$\hat{b}=\frac{\sum_{i=1}^{n}x_i\,y_i-n\,\overline{x}\,\overline{y}}{\sum_{i=1}^{n}x_i^2-n(\overline{x})^2}=\frac{L_{xy}}{L_{xx}}$$

$$\hat{a}=\overline{y}-\hat{b}\,\overline{x}$$

例2.1 为了研究大豆脂肪含量 x 和蛋白质含量 y 的关系, 测定了九种大豆品种籽粒内的脂肪含量和蛋白质含量, 得到如下数据

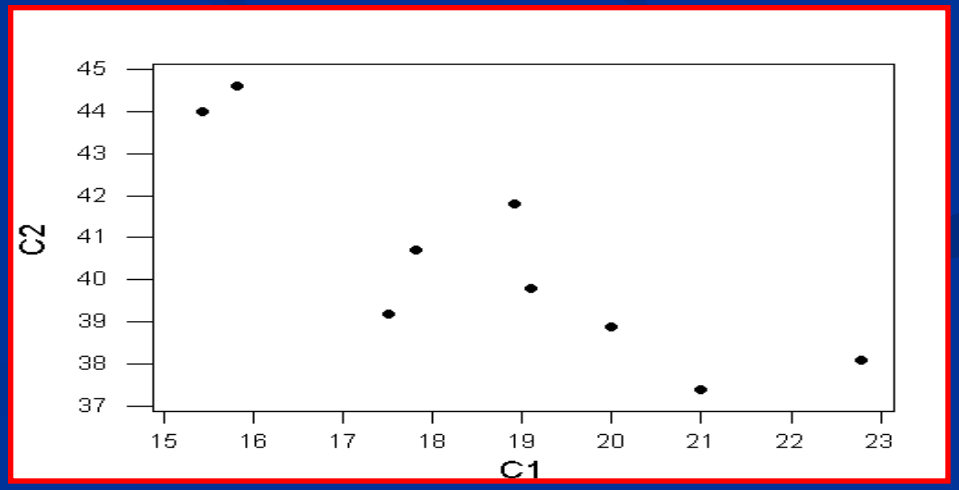
编号	1	2	3	4	5	6	7	8	9
脂肪含量%	15.4	17.5	18.9	20.0	21.0	22.8	15.8	17.8	19.1
蛋白质含量%	44.0	39.2	41.8	38.9	37.4	38.1	44.6	40.7	39.8

试求出 y 与 x 的关系。

解: （1）散点图、相关性

（2）建立回归模型、回归计算

（3）回归效果检验



$$\overline{x}=\frac{168.3}{9}=18.7;\quad \overline{y}=\frac{364.5}{9}=40.5$$

$$L_{xy}=\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}=6775.02-9\times18.7\times40.5=-41.13$$

$$L_{xx}=\sum_{i=1}^{n}x_i^2-n(\overline{x})^2=3192.75-9\times18.7^2=45.54$$

$$L_{yy}=\sum_{i=1}^{n}y_i^2-n(\overline{y})^2=14813.2-9\times40.5^2=50.95$$

$$\hat{b}=\frac{L_{xy}}{L_{xx}}\approx-0.9032\qquad \hat{a}=\overline{y}-\hat{b}\,\overline{x}=57.3891$$

故所求的回归方程为：

$$y=-0.9032x+57.3891$$

(4) 线性回归模型的评价

回归效果的好坏取决于回归方程的显著性检验。

原因： 由样本得到的 Y 关于 X 的线性回归方程

$$\hat{Y} = \hat{a} + \hat{b}X = \overline{Y} + \hat{b}(X - \overline{X})$$

是一条经过点 $(\overline{X}, \overline{Y})$，且斜率为 $\hat{b}$ 的直线。

由于任何杂乱无序，且无任何相关关系的散点都可以借助于直线或曲线去年拟合，但这种拟合是否有意义呢？因此，需要对回归的实际效果进行检验。

(5) 显著性检验的方法

① 相关系数法

$$r=\frac{\sum_{i=1}^{n}x_i y_i-n\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)\left(\sum_{i=1}^{n}y_i^2-n\overline{y}^2\right)}}$$

相关系数 r 越接近于1, 表明变量 y 与 x 线性相关程度越高, 但 r 究竟多大时, 就能说明变量 y 与 x 之间存在线性相关性, 从而用线性函数去拟合才算是合理的, 因此, 必须对相关程度进行显著性检验。

注: 利用相关系数检验回归的效果通常与所给的样本容量密切相关, 往往只有当样本容量较大时, 才能得出真正具有实际意义的回归方程。

表 1　简单相关系数的临界值表

$n-2$	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	1 000	0.062	0.081

例如：例题2.1 回归方程的有效性

$$R=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}\approx -0.8539 \qquad dof=n-2=9-2=7$$

查相关系数临界值表 $R_{0.01}=0.7977$

因为 $|R|>R_{0.01}$

所以回归方程在 $\alpha=0.01$ 的检验水平下有统计意义。

即可以认为大豆的蛋白质含量与脂肪含量有线性相关性。

② 方差分析法

设 (X_1，Y_1), (X_2，Y_2)，…，(Xn, Yn) 是一个容量为n的样本。

由样本得到的回归方程为：$\hat{Y} = \hat{a} + \hat{b}X$

且对应于自变量 X_i (I =1, 2, …， n)的回归估值Y_i，记为：

$$\hat{Y}_i = \hat{a} + \hat{b}X_i$$

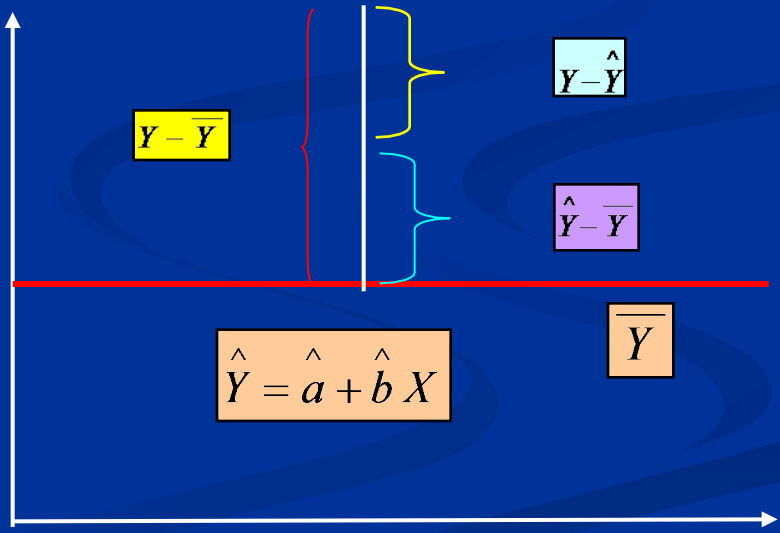
则样本观察值与均值的差：

$$Y_i - \overline{Y}$$

可分解为二部份

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

几何解释



其中, ($\hat{Y}_i - \overline{Y}$) 是由回归方程决定的，称为回归离差，而 ($Y_i - \hat{Y}_i$) 是由随机因素造成的，称为残差或误差。

平方和分解公式

可以证明：

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

SS_T总离差平方和

SS_R回归平方和

SS_e残差平方和

(i) 平方和

离差平方和

试验样本数据 $y_1, y_2, \ldots, y_n$ 之间存在差异是必然的，不可避免人。这种误差可由其试验数据与其平均值的偏差平方和加以刻划。

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = L_{yy}$$

回归平方和

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 L_{xx} = \hat{\beta}_1 L_{xy}$$

差异（偏差、波动）的产生是由两个方面的因素引起的：一是可控部分，这是由于控制变量x的变化而导致响应变量y发生变化,这种变化可用其回归估计值与其平均值的偏差平方和加以刻划。称为回归平方和.

残差（误差）平方和

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

二是随机部分，这是客观存在的,是不以人的意志为转移的波动。这类随机偏差可用样本数据与回归估计值的偏差平方和加以刻划。称为残差平方和.

三者之间的关系 $SS_T = SS_R + SS_e$

平方和分解公式

(ii) 自由度

总离差平方和 SS_T 的自由度 $df_T = n-1$

回归平方和 SS_R 的自由度 $df_R = 1$

一个回归变量

残差平方和 SS_e 的自由度 $df_e = n-2$

三者之间的关系 $df_T = df_R + df_e$

(iii) 均方

$$MS_T = \frac{SS_R}{df_R}, \quad MS_e = \frac{SS_e}{df_e}$$

方差分析表

一次线性回归方差分析表

差异源	SS 平方和	*df* 自由度	MS 均方	F 统计量	Fα 临界值	显著性
回归	SS_R	$df_T=1$	$MS_R= SS_R/ df_R = SS_R$	$F=MS_R/MS_e$	$F_\alpha=F(df_R, df_e)=F(1, n-2)$	* 显著
误差	SS_e	$df_e=n-2$	$MS_e= SS_e/ df_e = SS_e/ n-1$			**高度显著
总和	SS_T	$df_T=n-1$			$\alpha=0.05, 0.01$	

注：当线性回归方程显著时，说明 y 对 x 的变化主要是由于x 的变化所造成的；而当线性回归方程不显著时，说明 y 对 x 的变化除受 x 的变化影响外，还受到随机因素的影响.

③ 假设检验法(F — 检验法)

目的：检验自变量与所有因变量之间的线性关系是否显著。

步骤：

（一） 提出假设

原假设 $H_0: \beta = 0$ 接受此假设，意味着认为回归变量之间无线性关系。

备择假设 $H_1: \beta \neq 0$ 接受此假设，意味着认为回归变量之间有线性关系。

（二）构造检验统计量

残差平方和在回归平方和中所占的比重

① F 统计量

$$F = \frac{MS_R}{MS_e} = \frac{SS_R / dof_R}{SS_e / dof_e} = \frac{SS_R / 1}{SS_e / (n-2)}$$

可以证明：
当 H_0 成立时
$F \sim F(1, n-2)$

② 确定显著性水平

对给定的检验水平 α，查 F 分布临界值表，得相应的临界值

$$F_\alpha(1, n-1)$$

③ 统计推断

若 $F > F_\alpha(1, n-1)$ 称线性回归效果显著

也即认为变量 x 与 y 之间存在显著的线性关系。

若 $F < F_\alpha(1, n-1)$ 称线性回归效果不显著

也即认为变量 x 与 y 之间不存在显著的线性关系。

例如：例题2.1 回归方程的方差分析

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = L_{yy} = 50.95 \qquad f_T = 9-1 = 8$$

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 L_{xx} = \hat{\beta}_1 L_{xy} \qquad f_R = 1$$

$$= (-41.13)\times(-0.9032) = 37.15$$

$$SS_E = SS_T - SS_R \qquad f_e = f_T - f_R$$

$$= 50.95 - 37.15 = 13.80 \qquad = 8-1 = 7$$

方差分析表

差异来源	SS 平方和	*df* 自由度	MS 均方	F 统计量	Fα 临界值	显著性
回归	37. 15	1	37. 15	18.84	$F_{0.01(1,7)}=12.25$ $F_{0.05(1,7)}=5.59$	** 高度显著
误差	13.80	7	1.97			
总和	50.95	8				